# Beyond the Transfer-and-Merge Wordnet Construction: plWordNet and a Comparison with WordNet

**Marek Maziarz & Maciej Piasecki & Ewa Rudnicka**
Institute of Informatics
Wrocław University of Technology
Poland
mawroc@gmail.com
maciej.piasecki@pwr.wroc.pl
ewa.rudnicka78@gmail.com

**Stan Szpakowicz**
Institute of Computer Science
Polish Academy of Sciences, Poland
&
School of Electrical Engineering
and Computer Science
University of Ottawa, Canada
szpak@eecs.uottawa.ca

## Abstract

Wordnets are lexico-semantic resources essential in many NLP tasks. Princeton WordNet is the most widely known, and the most influential, among them. Wordnets for languages other than English tend to adopt unquestioningly WordNet's structure and its net of lexicalised concepts. We discuss a large wordnet constructed independently of WordNet, upon a model with a small yet significant difference. A mapping onto WordNet is under way; the large portions already linked open up a unique perspective on the comparison of similar but not fully compatible lexical resources. We also try to characterise numerically a wordnet's aptitude for NLP applications.

## 1 Introduction

It is hard to imagine NLP without lexico-semantic resources. The Princeton WordNet (PWN) is a powerful case in point: we have come to rely on it even in "hard-core" statistical methods of processing English texts. Wordnets for other languages, which soon followed PWN,[1] have usually been built by the transfer-and-merge method: the structure of PWN is copied over to the target language, the lexical material is translated, and the inevitable differences in language typology and cultural background are a matter of post-processing.[2]

The transfer-and-merge construction allows high compatibility between PWN and the target wordnet, so also between any wordnets built in the same way. Multi-lingual NLP work benefits from dependable interlingual relations – ensured if one uses a wordnet with PWN's structure. PWN's semantic relations are undoubtedly of general utility, but they do exhibit certain "English bias", and that – combined with the anglocentric network of concept underlying PWN's synsets – is a downside of the translation method of building a new wordnet.[3] The result need not be an accurate reflection of the lexico-semantic system of the target language.

The translation method has another advantage: it is rather affordable, because PWN is now very complete and quite stable. To start the construction of a wordnet without looking to PWN may seem a little foolhardy, but it offers certain intriguing benefits. This paper looks at one of such independent projects, a wordnet for Polish.

The plWordNet project aims to construct a large lexical resource (comparable in size to the largest existing wordnets, including PWN), based on few but precise principles and definitions. The goal is to achieve a faithful description of Polish while enabling compatibility with PWN (and by corollary with many wordnets), and yet avoid any semantic influences due to the transfer of the net of lexicalised concepts from PWN.[4] The work is semi-automatic and corpus-based. Linguists make final decisions, but supporting tools supply most of the raw material for those decisions, and naturally

[1] See `www.globalwordnet.org/gwa/wordnet_table.html` for an up-to-date list.

[2] Such differences are non-trivial even within the same language family – for example, Germanic, Romance or Slavic – and become highly significant as one moves further away from Indo-European languages.

[3] The term "concept" in this paper denotes objects which can be expressed by words. This deliberately skirts all the philosophical, cognitive and semantic issues, better left for another occasion.

[4] It would be impossible to avoid PWN's architectural influences. It is a model all wordnet creators aspire to.

*Proceedings of Recent Advances in Natural Language Processing*, pages 443–452,
Hissar, Bulgaria, 7-13 September 2013.

keep track of all aspects of the growing network.

No appropriately large machine-tractable thesaurus of Polish was available to jump-start the project. The construction has been based predominantly on the exploration of large corpora, with some help from traditional dictionaries. This required precise guidelines for linguists to facilitate the consistency of decisions and definitions – focused on linguistic data analysis and well anchored in the linguistic tradition. All information was fed into a steadily growing wordnet.

Today, plWordNet is large and mature enough to allow a wide-ranging observations. We analyse the consequences of the underlying wordnet model, the principles adopted, and the construction process. We take a varied perspective, including a multi-faceted comparison with PWN.

## 2 The structure of plWordNet

### 2.1 Constitutive relations

"A wordnet is a collection of synsets linked by semantic relations." This must be the most common quick take on wordnets in the literature. A synset is a set of synonyms which represent the same lexicalised concept, while synonyms are members of the same synset: this introduces a troubling circularity. An elaborate theory of synonymy could be a way of breaking the circle, but no such theory is operational enough in the sense of allowing precise guidelines for wordnet editors. This problem was discussed in (Derwojedowa et al., 2008; Piasecki et al., 2009; Maziarz et al., 2013).

Relations between synsets are often assumed to link concepts, and are fittingly described as conceptual relations. Their names, however, come up mainly in lexical semantics, where one considers hypernymy, meronymy etc. not between concepts but rather between words or lexical units (LUs).[5] Substitution tests usually proposed for synset relations refer to pairs of LUs (Vossen, 2002). Relations between LUs are relatively rare in PWN and in wordnets based on it, but antonymy, for example, never holds between synsets.

Neither concepts nor synsets occur directly in texts. LUs and their contexts of use *do* – and thus can be recognised, analysed and compared in corpora. This observation had led to a model of plWordNet different from that adopted by PWN: the basic building block is the LU, and semantic relations hold between LUs. A definition of a lexico-semantic relation includes a substitution test obligatorily applied by wordnet editors whenever a relation instance is added.[6]

The synset is a *secondary* notion. Synsets certainly appear in plWordNet, but they are defined via LUs. The cornerstone of this definitional machinery is a set of lexico-semantic *constitutive relations*, which contains in particular hypernymy, hyponymy, holonymy and meronymy. A relation is considered constitutive if its instances are frequent enough and frequently shared by groups of LUs.[7] It is also important that constitutive relations be established in linguistics (so wordnet builders feel comfortable around them) and accepted in the wordnet tradition (so compatibility among wordnets is easy to accomplish).

A synset is a group of LUs which share all constitutive relations; plWordNet software determines such groupings automatically. Thus, if relation $R$ is noted as linking synsets $S_1$ and $S_2$, it links every pair of LUs $s_1 \in S_1$ and $s_2 \in S_2$. An instance of a synset relation is naturally interpreted as an abbreviation for a set of LU relation instances.

It seems harder to recognise synonymy than LU pairs linked by constitutive relations. Relation instances are identified primarily via language data analysis (section 2.2). Avoiding the often troublesome synonymy is one of the facets of the *minimal commitment* principle which underlies the construction of plWordNet: make as few assumptions as possible. If no theory of meaning needs to be constantly invoked, and few intuitions about meaning variations are necessary, the construction process becomes "agnostic" about schools of linguistic thought. That is perhaps an opportunity: more applications are possible if fewer theoretical restrictions are imposed on a wordnet.

The relation set in plWordNet (Maziarz et al., 2011a; Maziarz et al., 2011b; Maziarz et al., 2012) elaborates on relations in PWN, EuroWordNet (Vossen, 2002) and GermaNet.[8] In addition to the expected (hyponymy, meronymy, antonymy, cause, instance for proper names, entailment – all adjusted to the reality of Polish), some relations account for the rich inflection and highly productive derivation of Polish. Assorted examples:

[5] A lexical unit is a lemma *and* its sense.

[6] An instantiated test is automatically presented by the editor-supporting software. As a tiny example, the test «if X is a Y, then "X" is a hyponym of "Y"» can be used to determine that in PWN *tiger 2* is a hyponym of *big cat 1*.

[7] As an example, antonymy is seldom shared, while it is common for several LUs to share a hypernym.

[8] `www.sfs.uni-tuebingen.de/GermaNet/`

INHABITANT (*góral* 'highlander' – *góry* 'highlands'); INCHOATIVITY (*zapalić się*$_{perf}$ 'ignite' – *palić się*$_{imperf}$ 'burn'); GRADATION (*gorący* 'hot' – *ciepły* 'warm' – *ciepławy* 'warmish', *letni* 'lukewarm' etc.); MODIFIER (*piwny* 'hazel' – *oko* 'eye'); PROCESS (*chamieć* 'roughen' – *cham* 'boor'); STATE (*panować* 'rule' – *władca* 'ruler'); AGENT (*spawacz* 'welder' – *spawać* 'weld'); INSTRUMENT (*nadajnik* 'transmitter' – *nadawać* 'transmit'); DIMINUTIVE (*córeczka* 'little daughter' – *córka* 'daughter').[9]

## 2.2 The construction process

Wordnet construction is rather like writing a dictionary (Fellbaum, 1998; Broda et al., 2012b). Lexicography distinguishes four phases: data collection, selection, analysis and presentation (Svensén, 2009). In the plWordNet project, language technologies support all four phases. Professional linguists under the supervision of senior coordinators work with *WordnetLoom*, a Web application. This graph-based wordnet editor allows visual browsing and concurrent editing. Many semi-automatic tools are integrated into *WordnetLoom*.

In the data collection phase, a large corpus is essential (Wynne, 2005). A multi-source corpus with 1.8 billion tokens, the foundation of plWordNet's systematic growth, supports the other phases of plWordNet's construction. The collected texts have been tagged by the morphological analyser *Morfeusz* (Woliński, 2006) and the TaKIPI tagger (Piasecki, 2007).[10]

In the data selection phase, the most frequent lemmas are chosen (plWordNet, 2012) and presented to the editors by *WordnetLoom*. The editors can also browse the plWordNet corpus using the Poliqarp interface (Janus and Przepiórkowski, 2005). To avoid time-consuming queries on the corpus, the process employs a word-sense disambiguation algorithm (Broda et al., 2010); it selects up to 10 examples of word usage, representing different meanings.[11] Finally, editing is supported by *WordnetWeaver* (Piasecki et al., 2009), a system which suggests several places where best to link a given lemma in the lexico-semantic net. Its hints usually yield new distinguished senses. The corpus browser, usage examples and *WordnetWeaver* enable increasingly complex language processing: from simple queries in the plWordNet corpus, through the presentation of a small list of disambiguated usage examples, to highly sophisticated lemma-placement suggestions.

In the data analysis phase, the editors answer a few central questions:

- whether a given lemma is correct in Polish (e.g., tagger mistakes are weeded out);
- how many LUs should be distinguished – whether all existing senses appear in usage examples or *WordnetWeaver*'s suggestions;
- how to describe a given LU by plWordNet relations – what relation types should be used.

Apart from primary sources and automated tools, the editors are encouraged to look up words and their descriptions in the available Polish dictionaries, thesauri, encyclopaedias, lexicons, and on the Web. At the end, the new lemma and all its LUs, or senses, are integrated with plWordNet and displayed in *WordnetLoom*.

Intuition matters despite even the strictest definitions and tests, so one cannot expect two linguists to come up with the same wordnet structure. In corpus-building it is feasible to have two people edit the same portion and adjudicate the effect, but wordnet development is a more complicated matter. That is why we have a three-step procedure: (i) wordnet editing by a linguist, (ii) wordnet verification by a coordinator (a senior linguist), and (iii) wordnet revision, again by a linguist. Full verification would be too costly, so it is done on (large) samples of the editors' work. A coordinator corrects errors, adjust the wordnet editor's guidelines,[12] and initiates revision during which systematic errors are corrected and the wordnet undergoes synset-specific modification.[13]

There also is a unique opportunity to verify the content of plWordNet meticulously: a mapping of its synsets onto PWN. That process sees every LU in plWordNet re-examined by a separate team of linguists. Section 4 explains in detail.

[9] MODIFIER is a syntagmatic relation. Its inclusion in plWordNet (rather like in Mel'čuk's Sense-Text Model) can add a lot of links, but we apply it in moderation.

[10] The corpus consists of 250 million tokens in the ICS PAS Corpus (Przepiórkowski, 2004); 113m tokens of news items (Weiss, 2008); ≈80m tokens in a corpus made of Polish *Wikipedia* (Wikipedia, 2010); an annotated corpus KPWr with ≈0.5m tokens (Broda et al., 2012a); ≈60m tokens of shorthand notes from the Polish parliament; and ≈1.2 billion tokens collected from the Internet.

[11] Usage examples, welcome by the editors, help them distinguish senses (Broda et al., 2012b).

[12] That is a 120-page document at present.

[13] All in all, an experienced editor, assisted by *WordnetLoom*, can increase plWordNet by up to 2000 LUs a month.

| wordnet | synsets | lemmas | LUs | avs |
|---|---|---|---|---|
| PWN | 117659 | 155593 | 206978 | 1.76 |
| *plWN* | 116323 | 106438 | 160100 | 1.37 |
| GermaNet | 74612 | 89819 | 99523 | 1.33 |

Table 1: The count of synsets, lemmas and LUs, and average synset size **avs**, in PWN 3.1, plWordNet 2.0 (*plWN*) and GermaNet 8.0.

| POS | synsets | lemmas | LUs | avs |
|---|---|---|---|---|
| N-PWN | 82115 | 117798 | 146347 | 1.78 |
| N-*plWN* | 80037 | 77662 | 109967 | 1.37 |
| V-PWN | 13767 | 11529 | 25047 | 1.81 |
| V-*plWN* | 21726 | 17486 | 31980 | 1.47 |
| A-PWN | 18156 | 21785 | 30004 | 1.65 |
| A-*plWN* | 14560 | 11290 | 18153 | 1.25 |

Table 2: The count of Noun/Verb/Adjective synsets, lemmas and LUs, and average synset size **avs**, in PWN 3.1 and plWordNet 2.0 (*plWN*).

### 2.3 The effects

A wordnet ought to be large to be really useful. Its coverage matters a lot to potential applications. Intuitively, the higher the coverage, the more information can be acquired from the resource. The size of plWordNet approaches that of PWN, a first for a resource not built by the transfer method. A comparison may not be foolproof given the different language typologies and plWordNet's choice of the lexical unit as a basic element, but it is quite instructive nonetheless.

#### 2.3.1 Size in numbers

Tables 1-2 present the statistics of the three largest manually constructed wordnets: Princeton WordNet 3.1, plWordNet 2.0 and GermaNet. PWN outstrips plWordNet when it comes to the number of lemmas and lexical units (word-sense pairs). Table 2 gives the precise counts of nouns, verbs and adjectives in PWN and plWordNet. The latter has more verbs, but fewer nouns and adjectives.

#### 2.3.2 Lexical coverage

The size of a wordnet can be contrasted with a frequency list from a large corpus. Such a measure of coverage sheds a light on the usability of a resource. A count was made of how many PWN lemmas appear in the text of English Wikipedia and how many plWordNet lemmas show up in the corpus described in section 2.2. The corpus sizes are comparable: *Wikipedia.en* has ≈1.2 billion words, the plWordNet corpus ≈1.4 billion words.[14] Table 3 shows percentages of wordnet noun lemmas by frequency bins (≥ 1000, 500, 200, 100, 50 occurrences). List of lemmas within particular frequencies are created from corpora, and then the presence of each of those lemmas in plWordNet or PWN is checked. The fast decreasing tails suggest that both wordnets more willingly absorb frequent lemmas than lemmas with lower frequencies. The plWordNet counts are higher simply because the same corpus underlies the frequency list and the vocabulary of plWordNet. The highest coverage ratio (≥ 1000) is much less than 100% because plWordNet contains almost no proper names.[15]

| FRC | ≥1000 | ≥500 | ≥200 | ≥100 | ≥50 |
|---|---|---|---|---|---|
| PWN | 0.383 | 0.280 | 0.170 | 0.107 | 0.064 |
| *plWN* | 0.535 | 0.456 | 0.350 | 0.277 | 0.210 |

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

#### 2.3.3 Polysemy

Table 4 shows the statistics of polysemy. Average polysemy is calculated by dividing the count of LUs by the count of lemmas. The column 'poly.' lists average polysemy for polysemous lemmas, the column '+mono.' gives the polysemy statistics for polysemous and monosemous lemmas together, the last column presents the ratio of polysemous lemmas to all lemmas. Nouns and adjectives are more polysemous in plWordNet than in PWN, verbs – conversely. This ought to be considered in the light of the part-of-speech statistics in Table 2 and the measure of corpus coverage in Table 3.

There are more nouns and adjectives in PWN, and since both wordnets tend to absorb high-frequency lemmas first, the polysemy in PWN must be lower. The paradox can be explained thus: the larger a wordnet, the higher the number of monosemous lemmas it contains, because more frequent lemmas are more polysemous. On the other hand, there are more monose-

[14] The corpus, with different genres and styles, is large enough to draw conclusions about coverage in applications.

[15] A large gazetteer with many semantic categories is ready to be incorporated into the wordnet (NELexicon, 2013).

| polysemy | poly. | +mono. | ratio |
|---|---|---|---|
| PWN - nouns | 2.38 | 1.24 | 0.18 |
| *plWN* - nouns | 2.57 | 1.42 | 0.26 |
| PWN - verbs | 2.93 | 2.17 | 0.60 |
| *plWN* - verbs | 3.00 | 1.83 | 0.41 |
| PWN - adjectives | 2.14 | 1.38 | 0.32 |
| *plWN* - adjectives | 2.59 | 1.61 | 0.38 |

Table 4: Average polysemy in PWN 3.1 and plWordNet 2.0 (*plWN*); poly. = only polysemous lemmas, +mono. = all lemmas, ratio = % of polysemous lemmas).

mous verb lemmas in plWordNet than in PWN. This puzzling difference between the ratio for polysemous verbs (2.93 vs. 3.00) and for all verb lemmas (2.17 vs. 1.83) can be explained if one assumes that in plWordNet polysemous verbs have statistically more fine-grained distinctions.

## 3 Indicators for WordNet 3.1 and plWordNet 2.0

### 3.1 Synset size

A relatively strict definition of synonymy and synsets adopted in plWordNet may be expected to lead to fewer lexical units per synset than in PWN. Column **avs** in Table 1 confirms: the average synset size in LUs is 1.37 and 1.76 respectively. Table 2 shows the averages per part of speech – the same overall effect. In general, plWordNet synsets are around 0.4 LU smaller than those in PWN. Statistics per domain, not shown here, also support this finding. The only larger difference occurs in the domain *animal*, probably because PWN synsets systematically include Latin names of species. For example, PWN has {dog 1, domestic dog 1, Canis familiaris 1} ‘a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds’. The equivalent plWordNet synset, linked by inter-lingual synonymy, is {pies 2} – just one common noun.

### 3.2 Relation density

Relation density comparison for PWN 3.1 and plWordNet 2.0 in Table 5 shows the average number of relations per synset.[16] The density is higher in plWordNet for nouns and verbs (+0.5 and +0.8 relation, respectively), lower for adjectives (-0.9). The total density is higher in plWordNet: on average, every other Polish synset has one synset relation instance more than PWN. The net is denser, a fact which can be explained like this: plWordNet has a stricter definition of synonymy, so there are more smaller synsets and thus the system needs more differentiating relations (and having more relations creates a feedback loop with a magnifying effect).

| POS | PWN | plWordNet |
|---|---|---|
| nouns | 3.54 | 3.99 |
| verbs | 2.21 | 3.06 |
| adjectives | 2.43 | 1.56 |
| total | 3.11* | 3.51 |

Table 5: Synset relation density in PWN 3.1 and in plWordNet 2.0 with regard to part of speech [*adverbs included].

The mapping of plWordNet onto PWN, described in detail in section 4, makes it possible to collate synsets from both wordnets linked by interlingual synonymy. It is interesting to see how relation density looks for corresponding synsets. Calculations have been run for all domains selected for mapping and described in Section 4 – see Table 6. For every plWordNet synset with interlingual synonymy, the count includes all relation instances to and from that synset, except obligatory inverse relations. The only outliers are the domains *body* and *location*: PWN has a higher density, even though Polish noun synsets have on average more relations than English noun synsets.

Now, locations and body parts are special vocabulary with many instances of meronymy. In plWordNet, meronymy suffices to link a new LUs to the net. In PWN, the most welcome relation for nouns is hyponymy. For example, {dłoń 1, ręka 3} ‘hand’ is a meronym of {ręka 1} ‘arm’, while its English I-synonym {hand 1, manus 1, mitt 1, paw 2} ‘the (prehensile) extremity of the superior limb’ is not only a meronym of {arm 1} ‘a human limb’, but also a hyponym of {extremity 5} ‘that part of a limb that is farthest from the torso’. Hyponymy is absent from plWordNet for synsets defined more naturally by part/whole semantics.[17]

[16] The count excludes obligatory inverse relations, usually counted in other publications (Tenenbaum, 2005, Table 2).

[17] Our policy is to avoid redundancy as much as possible.

| **POS** | **PWN** | **plWordNet** |
|---|---|---|
| **noun domains** | **PWN** | **plWordNet** |
| *artifact* | 3.90 | 4.83 |
| *body* | 8.06 | 6.70 |
| *communication* | 4.15 | 4.33 |
| *food* | 4.70 | 4.49 |
| *location* | 14.70 | 5.71 |
| *person* | 3.94 | 3.94 |
| *time* | 6.39 | 6.59 |

Table 6: Synset relation density in PWN 3.1 and in plWordNet 2.0 in selected domains.

| **path** | **avg.** | **std.** | **q1** | **q2** | **q3** | **max** |
|---|---|---|---|---|---|---|
| PWN *up* | 7.76 | 2.42 | 7 | 8 | 9 | 18 |
| *plWN up* | 5.71 | 3.33 | 4 | 6 | 8 | 21 |
| PWN *down* | 0.57 | 1.25 | 0 | 0 | 1 | 14 |
| *plWNdown* | 0.60 | 1.15 | 0 | 0 | 1 | 23 |

Table 7: Hypernymy path length for nouns in PWN 3.1 and plWordNet 2.0 (*plWN*). The headings: avg. = average, std. = standard deviation, q1, q2, 3 = quartiles; the minimum values are 0.

### 3.3 Hypernymy depth

A comparison of the average hypernymy depth in plWordNet and in PWN concerned noun synsets linked via inter-lingual synonymy and presumably located at the same or a very close level in the taxonomy. Next, the number of their intra-lingual relations *up* and *down* has been checked. The average hypernymy depth *up* is longer in PWN (7.76 relation) than in plWordNet (5.71). This is expected in view of the fact that PWN has a complex hyponymy structure above unique beginners and many of top synsets map straight to SUMO categories. plWordNet is mainly linguistically oriented, so there are very few SUMO categories in the hyponymy hierarchy (see Table 7).

The average hypernymy depth *down* is comparable: PWN 0.57, plWordNet 0.60. This is explained by the fact that the inter-lingual mapping was constructed bottom-up, thus at least half of the I-synonyms in both wordnets are leaves – the lowest nodes in the hierarchy.

## 4 Linking differently structured wordnets

A partial mapping of plWordNet onto PWN is ready (Rudnicka et al., 2012). A hierarchically arranged set of inter-lingual relations (I-relations) and a unique mapping procedure have been defined. The set was inspired by equivalence relations in EuroWordNet (Vossen, 2002) and by intra-lingual relations in plWordNet (Maziarz et al., 2011a). I-relations, complete with effective substitution tests, are considered in a strict order: I-synonymy, I-inter-register synonymy,[18] I-near-synonymy, I-hyponymy, I-hypernymy, I-meronymy, I-holonymy. The mapping procedure, working at the level of synsets, is based on a correspondence in meaning and position in the two wordnets' structures. There are three stages: recognize the sense of a source-language synset, find a target-language synset, and link the two synsets with one of the I-relations. Editors are supported by *WordnetLoom* (section 2.2) and by an automatic prompt system. They can also consult mono- and bilingual dictionaries.

The mapping is systematically verified. For the majority of the inter-lingual links entered thus far, a coordinator examines the source and target synsets' LUs and the type of the I-relation. The coordinator reviews any questionable link in *WordnetLoom* and either repairs it immediately or consults the editor in order to reach a consensus.

Besides the obvious advantage of building a bilingual wordnet, the mapping process enabled additional verification for plWordNet itself. The semantic domains selected for mapping were shared in such a way that one linguist constructed a particular plWordNet hypernymy branch and another linguist performed its mapping. This allowed re-editing the structure and content of plWordNet in case of mistakes. Linguists who did the mapping were encouraged to review critically the plWordNet side and introduce changes when they felt them necessary. The whole process was, naturally, regularly monitored by coordinators.

Table 8 shows the number of instances of I-relations in plWordNet 2.0 and in GermaNet 8.0, another partially manually constructed and mapped wordnet.[19] I-synonymy, a primary relation in both wordnets has a comparable number of instances. It is the most frequent relation in GermaNet, while in plWordNet it has been overtaken by I-hyponymy. The latter statistic can be explained by profound differences in the struc-

[18] Two LUs mean roughly the same but belong to different stylistic registers.

[19] We thank Verena Henrich for providing us with the relevant GermaNet data.

| Relation type | plWordNet 2.0 | GermaNet 8.0 |
|---|---|---|
| *I*-synonymy | 14240 | 15259 |
| *I*-hyponymy | 22873 | 1397 |
| *I*-hypernymy | 3329 | 760 |
| *I*-meronymy | 1732 | 126 |
| *I*-holonymy | 394 | 52 |
| *I*-near synonymy | 923 | 3389 |
| *I*-inter-register synonymy | 522 | — |

Table 8: Inter-lingual relation count (instances) in plWordNet and in GermaNet.

ture and content of plWordNet and PWN, discovered during mapping and discussed below. In GermaNet, I-hyponymy has quite few instances. On the other hand, the second largest relation in GermaNet is I-near synonymy.

There are lexico-semantic and lexico-grammatical differences between English and Polish: lexical and cultural gaps as well as different structuring of information, differences in the degree of gender lexicalisation and the frequency of marked forms such as diminutive or augmentative. Another type of contrasts is to do with the concept of synonymy and synsets, due mainly to the existence of "mixed" PWN synsets made up of neutral and marked, feminine and masculine, singular and plural, mass and count, and even hypernym and hyponym forms in the same synset. Additionally, hypernymy in plWordNet is strictly conjunctive (the meaning of a hyponym must comprise the meaning components of all its hypernyms), while PWN also allows disjunctive hypernymy (easily found in the glosses describing the meaning contribution of a given synset).[20] There are also differences in the use of more than one intra-lingual relation to code the same conceptual dependencies, various granularity of meaning description, and dictionary content mismatches.

Most, but not all, of these contrasts were accounted for by I-hyponymy: there were usually more lexically marked forms on the plWordNet side, while the larger, more general synsets were usually on the PWN side. It is another factor contributing to high hyponymy count in the overall statistics of relations.

Semantic domains selected for the first stage of mapping included *person*, *artefact*, *location*, *time*, *food* and *communication*. On average, the coverage of PWN domains amounts to approximately 50% of the respective plWordNet domain coverage, except for *location* where it is about 25%. That is mainly because the mapping went from plWordNet to PWN, but also because of the percentages of proper-name synsets. Proper-name synsets are rare in plWordNet – it was a deliberate decision – while they have a considerable share in PWN domains such as *person* and *location*.

The distribution of specific inter-lingual relations within the selected domains is as follows. For the most mapped domains – *person* and *location* – it mirrors the general distribution of I-relations (I-hyponymy slightly overtakes I-synonymy). For *artefact* and *communication* they are similar, while for *food* and *time* I-synonymy decidedly overtakes I-hyponymy. The high percentage of I-hyponymy in the *person* domain can be explained by the existence of many lexical and cultural gaps such as, for example, names of aristocratic titles or administrative functions, specific or even limited to one language community.

All in all, the set of inter-lingual relations and the mapping procedure developed for the purpose of mapping plWordNet, and the strategies of handling different types of mapping dilemmas, appear perfectly usable in linking other wordnets. The I-hyponymy links are now a clear sign of gaps which can be repaired in the further stages of the development of the networks. Mapping plWordNet to PWN also opens up the possibility of establishing links to other wordnets already linked to PWN.

## 5 Applications

Freely available for any purpose on a licence identical to the PWN licence, plWordNet has already proven its value in at least 16 research applications and in many publication which cite it.

The verb portion of plWordNet was used in semantic annotation in a corpus of referential gestures (Lis, 2012) and in a lexicon of semantic valency frames (Hajnicz, 2011; Hajnicz, 2012). In the latter, plWordNet domains were also used in algorithms of verb classification. In (Maciołek, 2010; Maciołek and Dobrowolski, 2013) plWordNet is used to extend a set of features for text mining from Web pages. In (Wróblewska et al., 2013)

[20] Glosses for all synsets are a relatively late addition to PWN. We have only recently begun to introduce them into plWordNet.

plWordNet was the basis for building a mapping between a lexicon and an ontology. Miłkowski (2010) included plWordNet in a set of dictionaries in his proofreading tool. There are applications of plWordNet in word-to-word similarity measures utilised in research on ontologies (Lula and Paliwoda-Pękosz, 2009) or in calculating text similarity (Siemiński, 2012). As a semantic lexicon, plWordNet has been useful in text classification (Maciołek, 2010), terminology extraction and clustering (Mykowiecka and Marciniak, 2012), automated extraction of opinion attribute lexicons from product descriptions (Wawer and Gołuchowski, 2012), named entity recognition, word-sense disambiguation, extraction of semantic relations (Gołuchowski and Przepiórkowski, 2012), temporal information (Jarzębowski and Przepiórkowski, 2012) and anaphora resolution.

Open Multilingual Wordnet (Bond, 2013) now includes plWordNet. It is referred to in other work 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).

The resource has attracted about 450 registered individual and institutional users (registration upon download is not mandatory). The plWordNet Web page and Web service have had tens of thousands of visitors (hundreds of thousands of searches). The intended use includes 70 commercial applications, and 50 scientific and educational applications (at all levels: university, high school and primary school). The declared topics of scientific applications include semantic word similarity calculation, multilingual word-sense disambiguation, text classification, knowledge base for recommender systems and information retrieval (e.g., wordnet-based query expansion, user modelling, personalisation and user profile), Question Answering, Information Extraction systems (including automated event extraction), Text Mining, Opinion Mining, parsing disambiguation, ontology-based systems (ontology construction, integration and mapping to a lexicon), comparative research on languages and wordnets, chatbot systems (as a lexicon), text similarity in processing legal texts, anti-plagiarism, contrastive/comparative studies (e.g., "Comparison of Polish, English and Swedish terms of motion and emotion, including analysis of metaphorical expressions." or "Conducting a cross-linguistic study on phonesthemes."), Affect Analysis (multilingual systems), humour analysis, development of Polish Link Grammar, and plWordNet as an object of analysis of complex networks.

Companies downloaded plWordNet for knowledge base management systems (e.g., automated conversion of text documents into a knowledge base), Business Intelligence, document similarity calculation, Polish website mapping and keyword tracking, online multilingual dictionary, search engine component development, translation inference support, analysis of public discourse, use as an additional bilingual dictionary in translation practice, Question Answering, text verification during editing, meta-data for publications, Polish dictionary and a basis for the development of bilingual dictionaries.

In education, plWordNet was named in many student projects in NLP, lectures on NLP, a course on Text mining for sociologists. It has been also utilised in teaching linguistics and even as an illustration of linguistic notions in education in primary and secondary schools.

## 6 Conclusions

The paper has discussed the construction of plWordNet, a national wordnet not adapted from Princeton WordNet by the transfer-and-merge method. The present contents of plWordNet are comparable in size to "The Mother of All WordNets", as well as in lexical coverage, hypernymy depth and relation density. The treatment of synonymy and synsets is an alternative to the usual model adopted in PWN and numerous other wordnets: synset membership depends only on constitutive relations between lexical units.

In its current mature stage of development, plWordNet is being mapped onto PWN. A unique mapping strategy aims at linking synsets based on the correspondence of meaning and position in the wordnet structure. The mapping process has revealed a number of contrasts between the two networks. They can be explained by lexico-grammatical differences between English and Polish, and the subtly different methodologies behind the construction of the two networks.

## Acknowledgment

Work financed by the European Union, Project POIG.01.01.02-14-013/09 in the European Innovative Economy Programme.

## References

Francis Bond. 2013. Open multilingual wordnet. Web page of the resource and project: `http://casta-net.jp/~kuribayashi/multi/`, May.

Lars Borin and Markus Forsberg. 2010. From the people's synonym dictionary to fuzzy synsets – first step. In *Proceedings of LREC 2010*.

Bartosz Broda, Marek Maziarz, and Maciej Piasecki. 2010. Evaluating LexCSD — a Weakly-Supervised Method on Improved Semantically Annotated Corpus in a Large Scale Experiment. In S. T. Wierzchoń M. A. Kłopotek, A. Przepiórkowski and K. Trojanowski, editors, *Proceedings of a Conference on Intelligent Information Systems*.

Bartosz Broda, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardyński. 2012a. KPWr: Towards a Free Corpus of Polish.

Bartosz Broda, Marek Maziarz, and Maciej Piasecki. 2012b. Tools for plWordNet Development. Presentation and Perspectives. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3647–3652, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors. 2012. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012*. European Language Resources Association (ELRA).

Magdalena Derwojedowa, Stanisław Szpakowicz, Magdalena Zawisławska, and Maciej Piasecki. 2008. Lexical Units as the Centrepiece of a Wordnet. In Mieczysław A. Kłopotek, Adam Przepiórkowski, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors, *Proceedings of the 16th International Conference Intelligent Information Systems*, Advances in Soft Computing, pages 351–358, Warsaw. Academic Publishing House EXIT.

Christiane Fellbaum. 1998. A Semantic Network of English: The Mother of All WordNets. *Computers and the Humanities*, 32:209–220.

Konrad Gołuchowski and Adam Przepiórkowski. 2012. Semantic role labelling without deep syntactic parsing. In Isahara and Kanzaki (Isahara and Kanzaki, 2012), pages 192–197.

Elżbieta Hajnicz. 2011. Grouping alternating schemata in semantic valence dictionary of polish verbs. In *Proceedings of Text, Speech and Dialogue*, volume 6836 of *Lecture Notes in Computer Science*, pages 155–162. Springer.

Elżbieta Hajnicz. 2012. Similarity-based method of detecting diathesis alternations in semantic valence dictionary of polish verbs. In *Proceedings of Security and Intelligent Information Systems*, volume 7053 of *Lecture Notes in Computer Science*, pages 345–358.

Hitoshi Isahara and Kyoko Kanzaki, editors. 2012. *Advances in Natural Language Processing: Proceedings of the 8th International Conference on NLP, JapTAL 2012, Kanazawa, Japan, October 22-24, 2012*, volume 7614 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag, Heidelberg.

Daniel Janus and Adam Przepiórkowski. 2005. Poliqarp 1.0: Some technical aspects of a linguistic search engine for large corpora. In *The proceedings of Practical Applications of Linguistic Corpora*.

Przemysław Jarzębowski and Adam Przepiórkowski. 2012. Temporal information extraction with cross-language projected data. In Isahara and Kanzaki (Isahara and Kanzaki, 2012), pages 198–209.

Krister Lindén and Lauri Carlson. 2010. Finnwordnet – wordnet på finska via översättning. *LexicoNorcdica*, 17.

Magdalena Lis. 2012. Polish multimodal corpus - a collection of referential gestures. In Calzolari et al. (Calzolari et al., 2012), pages 1108–1113.

Paweł Lula and Grażyna Paliwoda-Pękosz. 2009. PodobieŃstwo semantyczne w analizie danych przekrojowych. In Krzysztof Jajuga and Marek Walesiak, editors, *Taksonomia 16 Klasyfikacja i analiza danych — teoria i zastosowania*, Prace naukowe Uniwersytetu Ekonomicznego we Wrocławiu, pages 104–112. Uniwersytet Ekonomiczny we Wrocławiu.

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

Przemysław Maciołek. 2010. Is shallow semantic analysis really that shallow? a study on improving text classification performance. In *Proceedings of the 2010 International Multiconference on Computer Science and Information Technology (IMCSIT)*.

Marek Maziarz, Maciej Piasecki, Stan Szpakowicz, and Joanna Rabiega-Wiśniewska. 2011a. Semantic Relations among Nouns in Polish WordNet Grounded in Lexicographic and Semantic Tradition. *Cognitive Studies*, 11:161–181.

Marek Maziarz, Maciej Piasecki, Stan Szpakowicz, Joanna Rabiega-Wiśniewska, and Bożena Hojka. 2011b. Semantic Relations Between Verbs in Polish Wordnet 2.0. *Cognitive Studies*, 11:183–200.

Marek Maziarz, Stan 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*. DOI 10.1007/s10579-012-9209-9, 28 pages.

Verginica Barbu Mititelu. 2012. Adding morpho-semantic relations to the romanian wordnet. In *Proceedings of LREC 2012*.

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

Agnieszka Mykowiecka and Małgorzata Marciniak. 2012. Combining wordnet and morphosyntactic information in terminology clustering. In *Proceedings of COLING 2012: Technical Papers COLING 2012, Mumbai, December 2012.*, pages 1951–1962.

NELexicon. 2013. NELexicon: a gazetteer of proper names for Polish. `www.nlp.pwr.wroc.pl/en/tools-and-resources/nelexicon`.

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*.

Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. 2009. *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wrocławskiej. `www.plwordnet.pwr.wroc.pl/main/content/files/publications/A_Wordnet_from_the_Ground_Up.pdf`.

M. Piasecki. 2007. Polish Tagger TaKIPI: Rule Based Construction and Optimisation. *Task Quarterly*, 11(1–2):151–167.

plWordNet. 2012. Frequency List from plWorNet Corpus. `www.nlp.pwr.wroc.pl/pl/narzedzia-i-zasoby/lista-frekwencyjna`.

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

Ewa Rudnicka, Marek Maziarz, Maciej Piasecki, and Stanisław Szpakowicz Szpakowicz. 2012. A strategy of Mapping Polish WordNet onto Princeton WordNet. In *Proceedings of COLING 2012*. ACL.

Andrzej Siemiński. 2012. Fast algorithm for assessing semantic similarity of texts. *International Journal of Intelligent Information and Database Systems*, 6(5):495–512.

Bo Svensén. 2009. *A Handbook of Lexicography. The Theory and Practice of Dictionary-Making*. Cambridge University Press.

Mark Steyvers & Joshua B. Tenenbaum. 2005. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29:41–78.

Piek Vossen. 2002. EuroWordNet General Document Version 3. Technical report, Univ. of Amsterdam.

Aleksander Wawer and Konrad Gołuchowski. 2012. Expanding opinion attribute lexicons. In *Proceedings of Text, Speech and Dialogue, Brno 2012*, volume 7499 of *Lecture Notes in Computer Science*, pages 72–80. Springer.

Dawid Weiss. 2008. Korpus Rzeczpospolitej. [on-line] `www.cs.put.poznan.pl/dweiss/rzeczpospolita`. Corpus of text from the online edtion of Rzeczypospolita.

Wikipedia. 2010. Polish Wikipedia. online. `pl.wikipedia.org`, accessed in 2010.

Marcin Woliński. 2006. Morfeusz – a practical tool for the morphological analysis of Polish. In M.A. Kłopotek, S.T. Wierzchoń, and K. Trojanowski, editors, *Intelligent Information Processing and Web Mining: Proceedings of the International Conference IIS: IIPWM'06*, Advances in Soft Computing, pages 511–520, Berlin. Springer.

Anna Wróblewska, Grzegorz Protaziuk, Robert Bembenik, and Teresa Podsiadły-Marczykowska. 2013. Associations between texts and ontology. In *Intelligent Tools for Building a Scientific Information Platform*, volume 467 of *Studies in Computational Intelligence*, pages 305–321. Springer.

M. Wynne, editor. 2005. *Developing Linguistic Corpora: a Guide to Good Practice*. Oxbow Books, Oxford.

Ayesha Zafar, Afia Mahmood, Farhat Abdullah, Saira Zahid, Sarmad Hussain, and Asad Mustafa. 2012. Developing urdu wordnet using the merge approach. In *Proceedings of the Conference on Language and Technology*, pages 55–59.

Krešimir Šojat, Matea Srebačić, and Marko Tadić. 2012. Derivational and semantic relations of croatian verbs. *Journal of Language Modelling*, 0(1):111–142.