plWordNet 4.1 – a Linguistically Motivated, Corpus-based Bilingual Resource

Agnieszka Dziob, Maciej Piasecki and Ewa Rudnicka
G4.19 Research Group, Department of Computational Intelligence
Wrocław University of Technology, Wrocław, Poland
{agnieszka.dziob,maciej.piasecki,ewa.rudnicka}@pwr.edu.pl

Abstract

The paper presents the latest release of the Polish WordNet, namely plWordNet 4.1. The most significant developments since 3.0 version include new relations for nouns and verbs, mapping semantic role-relations from the valency lexicon Walenty onto the plWordNet structure and sense-level interlingual mapping. Several statistics are presented in order to illustrate the development and contemporary state of the wordnet.

1 Introduction

plWordNet (Pol. *Słowosieć*) is a very large wordnet of Polish, mapped to Princeton WordNet of English (Miller et al., 1990) and enriched with other links and annotations. Its development started back in 2005, and has been continued since then. In 2016, its complex, mature 3.0 version was presented in (Maziarz et al., 2016). It achieved very large size and coverage of words in Polish corpora. Thus, in our work we focused on increasing the density of the network of wordnet relations, revising the structure wherever necessary and adding new relations in order to improve the description of the lexical system of Polish and to meet the requirements of plWordNet's applications.

The goal of this paper is to present the latest 4.1 release of plWordNet, the result of a linguistically motivated expansion of 3.0 version. We will discuss new synset relations for nouns and verbs, as well as a new relation for lexical units, namely the semantic Collocation relation meant to facilitate the use of plWordNet in Word Sense Disambiguation. A new system of verb classes will be briefly recalled with the focus on its implementation in 4.1. We will also discuss the process of systematic assignment of aspect values, such as perfect, imperfect, and bi-aspectual, to every verbal lexical unit.

Alongside the development of plWordNet, works on its mapping to Princeton WordNet are carried out. The latest version includes the complete mapping of Polish and English noun synsets, the extended mapping of adjective and adverb synsets and the substantial mapping of verb synsets. Moreover, we have started the development of a system of equivalence relations for noun lexical units. Finally, we will present the results of mapping plWordNet lexical units onto the entries of the Polish valency lexicon *Walenty* and their partial manual verification.

2 Linguistic Motivation

Since its origin, plWordNet has been built around the idea of making lexical units[1] (henceforth, LUs) its basic building blocks, using linguistic lexico-semantic relations, and making the wordnet a faithful description of the Polish lexical system, see (Piasecki et al., 2009). This led to a corpus-based wordnet development process (Maziarz et al., 2016), synset composition based on sharing constitutive relations and features, and wordnet model based on the Minimum Commitment Principle, see (Maziarz et al., 2013b).

In plWordNet, the description of lexical meanings is primarily based on lexico-semantic relations directly originating from lexico-semantic relations known from lexicography. As synset relations are abbreviations for the fact of sharing relation between LUs – synset components – synset relation do not differ in their character from relations linking LUs. Glosses and usage examples are treated

[1] A lexical unit is technically defined as a triple: lemma, Part of Speech and sense identifier.

as secondary means of description. Definitions of particular relations directly refer to language data via substitution tests that are then used in the wordnet development.

In plWordNet 3.0 (Maziarz et al., 2016) LUs located in the lower parts of the hypernymy hierarchy were often described by only a few relation links, if not just one. Thus, their meaning descriptions were limited, especially those described by single hyponymy links connecting to the same hypernym. There was no meaning distinction between such LUs. Diversity and density of relation links is crucial for many applications of a wordnet, e.g. comparison of meanings, analysis of selectional preferences (McCarthy and Carroll, 2003, Hajnicz et al., 2016), Word Sense Disambiguation (Agirre and Rigau, 1996, Kędzia et al., 2015), texts semantic indexing (Scott and Matwin, 1998), query expansion in Information Retrieval (Voorhees, 1998, Varelas et al., 2005), or construction of topic descriptors for media monitoring (Johansson et al., 2012).

Taking the above into account, we have proposed an expansion of the plWordNet model by several new relations, described in the next section. In sum, we will have 33 types of synset relations (52 when counting subtypes) and 20 types of LU relations, i.e. not shared among LUs (56 including subtypes).

2.1 Nouns

The system of noun relations in plWordNet 4.1 is based on that of 3.0 release (Maziarz et al., 2011). It has recently been expanded with several new synset relations discovered when analysing instances of the *fuzzynymy* relation. During the many years of plWordNet development, *fuzzynymy* was used as a kind of notebook to record semantic associations that seemed prominent, but irregular from the point of view of the wordnet relation system. Still, not all fuzzynymy relations were renamed into other relations.

Definitional feature is a relation informing about an entity's intrinsic property which defines its membership to a given class of things or people e.g. {*rudzielec 1, marchewka 3, wiewióra 1*} '≈redhead' →{rudy} 'red', {*upał 1, skwar 1, żar 1, spiekota 1, spieka 1*} '≈heat'→{*upalnie 1, skwarno 1, skwarnie 1*} 'hot', {*abrakadabra 1, metafizyka 3, czarna magia 1*} 'double Dutch' →{*niezrozumiały 1*} 'unclear'. It is a relation between a noun synset and another noun synset or an adjective or adverb synset. This property rarely co-occurs with a given noun in the corpus, but it often appears in its lexical paraphrase.

Area of interest is a noun-noun relation that informs about an object or issue that is lexically constituted as a typical focus for this discipline or area, e.g. {*kardiologia 2*} 'cardiology'→{*układ krwionośny 1, krwiobieg 1, krwioobieg 1*} 'circulatory system'.

Origin is a relation linking a noun with a qualitative adjective derived from a noun denoting the country or culture of origin of the entity denoted by this noun e.g. {*zabaglione 1, zabajone 1, zabaione 1*} 'sabayon'→{*włoski 3, italiański 3, italski 3*} 'Italian', or, when there is no such adjective, with a noun denoting the origin of a given entity. It can be paraphrased as 'something that comes from a country or culture'.

Parameter is a noun-noun relation defined especially for the description of specialist vocabulary and represents a physical, measurable parameter characterising some phenomena, e.g. {*żyzność 1, urodzajność 1, żyzność gleby 1, plenność 1*} 'soil fertility'→{*gleba 1, grunt 3, podłoże 2*} 'ground'. Specialist vocabulary LUs are almost always found in the lower part of the hypernymy hierarchy and are described by few relations besides hyponymy.

Following our earlier positive experience in using a derivationally motivated Role of a hidden predicate relation linking noun LUs since plWordNet 2.0 (Maziarz et al., 2011), we propose to expand this relation to relations between synsets in which the semantic opposition is similar, but the linked synset elements are usually not derivationally associated.

Subject of hidden predicate is a relation between two noun synsets such that the first is a semantic subject of an implicit action intentionally and intrinsically related to an object represented by the second, e.g. {*pulmonolog 1, pneumonolog 2*} 'pulmonologist'→{*układ oddechowy*} 'respiratory system'.

Product|result of hidden predicate, in a similar way, associates two noun synsets such that the first represents a product or result of an implicit action or process done on an object or substance represented by the second, e.g. {*piwo 1, złoty trunek 1, złocisty trunek 1*} 'beer'→{*brzeczka piwna 1*} 'beer wort'.

Place of hidden predicate links two noun synsets where the first represents a place which is an obligatory, lexically constituted element of an action or process represented by an implicit predicate which is intrinsically related to the entity expressed by the second synset, e.g. {*klinika odwykowa*} 'rehab clinic'→{*nałogowiec 1, uzależniony 1*} 'addict'.

Although synset relations based on the hidden predicate scheme are mostly used for the description of specialist vocabulary, they are quite frequent, i.e. a couple of hundred instances on average, see Sec. 3.

2.2 Collocation

A large wordnet can be successfully used as a knowledge base for Word Sense Disambiguation, but the quality of the resulting system depends a lot on the richness of a network of connections between words from texts via their senses, especially between senses that are likely to co-occur in similar contexts (Leacock et al., 1998). Unfortunately, the coverage of such associations is limited by typical wordnet relations.

Following the above observation, we introduce a *collocation* relation for LUs that links lexical meanings, not words. In contrast to the *definitional feature* relation, which is based on a semantically motivated, paradigmatic feature, *collocation* follows the corpus supported language data and indicates frequent meaning co-occurrences. It can link two LUs of any of part of speech, if they co-occur often enough in corpora. So far we have added 16 979 instances of the collocation relation for all parts of speech to plWordNet (most of them for nouns: 7 838 instances).

Collocation relation was also used for priming selected meanings for words, as a kind of micro-glosses during psychological experiments on collecting emotive evaluations per LUs. For a selected subset of polysemous lemmas, we first drew one LU per lemma as subjects of the experiment. Next, for each selected LU we tried to choose among its possible frequent co-occurrences in such a way that the chosen collocation distinguish the given LU (word meaning) from all the other possible ones for a given lemma. Later, during the experiment, a lemma – representing an intended LU – was presented alongside the collocation to the informants, who were next interviewed about their reactions to several emotive aspects of the LU meaning. *Collocation* as defined and used by us has a pragmatic application: it links meanings, not words, according to their co-occurrence in corpora, while typical statistical analysis of corpora yields only word-form associations. We therefore regard this relation useful for word sense disambiguation. So far, the collocation relation, as it is proposed, has only a utilitarian character. However, we plan further research in this field.

2.3 Adjectives and Adverbs

The description of adjectives in plWordNet 3.0 was based on several synset relations, including *inter-register synonymy*, *hypernymy/hyponymy*, *gradation*, *modifier*, and *value (of the attribute)* (Maziarz et al., 2012). The synset relations were complemented by a set of LU relations, e.g. *predisposition* (with 4 subtypes), *role Adj-V* (7 subtypes), *antonymy* (complementary and gradable), *cross-categorial synonymy* to nouns (2 subtypes), *characteristic*, *markedness*, *role: material* or *state/feature* (derivationally motivated). We found this whole system of relations working well, so except for *definitional feature* proposed in Sec. 2.1 and *collocation* described in Sec. 3, we do not propose any changes to it. Instead, the coverage of several adjective relations was expanded.

Adverbs are treated similarly to adjectives in plWordNet 4.1. Their model in 3.0 version encompassed a set of synset relations almost identical to adjective relations – with the exclusion of *modifier* – and a set of LU relations including: *antonymy* (complementary and gradable), and *cross-categorial synonymy* to adjectives. Similarly to the adjective relations, we keep the adverb relations unchanged.

2.4 Verbs

Verbs in plWordNet 3.0 were organised in a sophisticated system of hierarchical semantic classes that influenced or even determined the verb relation structure. The classification encompassed 9 main classes and 4 auxiliary subclasses and was based on the proposal of (Laskowski, 1998), which has never been verified on large language data. This system made plWordNet 3.0 difficult to edit and led to criticism of the excessive proliferation of verb senses (Dziob and Piasecki, 2018b).

Dziob and Piasecki (2018a) proposed a much simpler verb classification for plWordNet consisting of just two main semantic verb classes, namely *static* and *dynamic* verbs. Only this division is reflected in definitions of several selected verb relations. In addition, for dynamic verbs five subclasses were proposed, namely: *distributive*, *accumulative*, *perdurative*, *delimitative* and *action* verbs, but without the obligatory influence on their relations. The decision about the verb class membership of a given LU is done with the help of semantic paraphrases, which simplifies the work of lexicographers and results in a description that is more comprehensible for users. (Dziob and Piasecki, 2018a) proposed a couple of new verb relations and modifications to several relations which we have adopted for plWordNet 4.1 and describe below.

We decided to leave two main subtypes of the *aspectuality* relation: *pure* and *secondary*, which express the basic semantic difference between aspectual pairs in Polish, (Dziob et al., 2017). Yet, since aspect has become a feature assigned to the verb, we have decided against further division of aspectuality and other relations, based only on aspect. Therefore, this system has become simpler. Otherwise, we have introduced a few new types and subtypes of verb relations: for backward relations (*preceding* and *presupposition*) subtypes without subject identity (e.g. *rozwieść się* 'to get divorced'←*małżeństwo* 'marriage') and four new main level relations, based on syntagmatic occurrences and also lexical definitions: *subject*, *object*, *circumstance* and *manner*, see (Dziob and Piasecki, 2018a). An important change is the possibility of linking verbs with adverbs, allowed since 3.1 version.

3 Structure

Since 3.0 version, we have expanded plWordNet both in terms of language material covered and the number of relation links, characterised briefly in this section and shown in Tab. 3. As main goals for the expansion to plWordNet 4.1 we identified: the newest Polish vocabulary (and meaning changes) in relation to the whole lexical system of Polish and specialist terminology (including multi-word expressions) from users' corpora. Concerning the first goal, this is a necessary process for preserving the quality of plWordNet as a comprehensive and up-to-date description of the Polish lexical system. Continuous development of the coverage of a wordnet is an obligatory aspect for the preservation of its quality.

The presence of specialist vocabulary, mostly terminology, in a general dictionary (a large wordnet is often perceived as a large dictionary) is disputable. However, plWordNet is mostly used as a basic language resource in processing, e.g. as the part of the CLARIN language technology infrastructure, and its content should reflect to some extent the vocabulary of texts being processed. As such, the addition of specialist terminology and vocabulary has been a corpus-driven effort.

The development of plWordNet follows the corpus-based wordnet development process proposed in Maziarz et al. (2013a). plWordNet Corpus v10 has been enlarged up to 4.2 billion segments in order to make it a better basis for the acquisition of new lemmas. New colloquial vocabulary was added from sources such as social media, blogs, also the most recent literature. Several much smaller specialist corpora from the CLARIN-PL users were also explored as the sources of language material.

3.1 Changes in Statistics

Since we suspected that adjective and adverb parts were less developed, we compared their content with the plWordNet Corpus. All missing adjectives and adverbs were added and the meanings of many of them were verified that resulted in expanding plWordNet by at least 2300 adjective lemmas (>8,500 adj. LUs) and 2000 adverb lemmas (>3,100 adv. LUs).

As the verb model had been changed and we knew that the coverage for verbs was lower than for other parts of speech, we put special emphasis on a large scale expansion of this sub-database and also on the verification and correction of the existing description of many verbs. More than 11,300 new verb LUs and

Elements	**Verbs**	**Nouns**	**Adv.**	**Adj.**	**All**	↑
plWN 3.0 Lemmas	17 398	126 746	5 719	27 041	**177 003**	–
plWN 3.0 Lexical Units	31 841	167 243	10 416	45 899	**255 733**	–
plWN 3.0 Synsets	21 669	123 985	8 080	39 204	**193 286**	–
plWN 4.1 Lemmas	20 430	134 674	8 042	29 349	**192 495**	8.7%
plWN 4.1 Lexical Units	43 701	178 167	14 088	54 410	**290 366**	13.5%
plWN 4.1 Synsets	32 102	133 747	11 295	47 035	**224 179**	16.0%

Table 1: Basic statistics of plWordNet 4.1 (`http://plwordnet.pwr.edu.pl`)

2,900 new verb lemmas were added. The new LUs were also added to lemmas already present in the 3.0 version to complete the description of their meanings. Manual verification and correction of LUs, synsets and relations was done for most of the already described verbs.

Specialist vocabulary was added in response to requirements of plWordNet applications (esp. in CLARIN) subsuming about 4,000 specialist LUs (mostly nouns, marked by *specialist* register), including many multi-words. The newest vocabulary acquired from plWordNet Corpus v10 was described by more than 1,500 new LUs of all parts of speech. Changes in the noun part are mostly the result of this process.

Tab. 3.1 presents statistics for the proposed noun relations. Because we have started to add new relations to the specialist vocabulary, *area* and (especially) *parameter* are not too frequent relations, but development of this vocabulary is ongoing. We use specialist corpora of CLARIN users and we integrate the vocabulary derived from them by means of these relations. We expect them to be useful especially for describing specialist vocabulary on the lowest levels of the wordnet hierarchy – at least this is the result of our experience up to now.

The second source of new lemmas and lexical units are users' diachronic corpora containing old vocabulary. We include these units in plWordNet only when we can confirm their use in texts, for example in freely available old literature. For this reason and because of the presence of modern vocabulary in our corpora that we write about in Sec. 3., the quantity of *inter-register synonymy* linking synsets with LUs of divergent registers is increasing (in 4.1 version it amounts to 12 223 instances for all parts of speech, of which most for nouns – 7 171). We expect that this process will advance.

3.2 Non-relational Elements and Verification

In 2017 a wordnet editor system called WordnetLoom 2.0 (Naskręt et al., 2018) was enriched with the ability to record comments concerning the correctness of a given LU and synset. Information that has been collected by this system is one of the inputs to the plWordNet verification process started by us. We use also data collected from the diagnostic tools (Piasecki et al., 2016).

The verification of plWordNet is performed on the two levels of LUs and synsets. Both are described by an additional *status* feature whose value is set by a lexicographer after each operation: *verified*, *partially processed*, *new*, *meaning*, *erroneous* and *not processed*, with the last one as a default value. When an editor spots a problem they can describe or comment on it. In this way, statistics concerning the frequency of errors made during the earlier stages of plWordNet development are collected. The most frequent errors are: too small number of meanings for a given lemma, but also too fine-grained granulation of meanings, and wrong stylistic register. The verification and corrective editing that has been performed since the publishing of the 3.0 version is focused on LUs now, as we assume that a *verified* synset must include only *verified* LUs. A fully correct synset must include LUs with proper descriptions, including their relations, and the synset must be described by proper synset relations (compatible with the LUs due to the synset definition assumed in plWordNet). So far 7,976 have been marked by the status *verified* and 5,677 *partially processed*, i.e. verified by a single editor and waiting for the confirmation by the second editor.

The description of LUs in plWordNet is systematically completed by glosses and use ex-

Rel. of hidden pr. (general)	855
Parameter	83
Origin	1324
Area	344
Definitional feature	660

Table 2: Statistics of new relations for nouns.

amples (both added on the level of LUs, not synsets). None of them are necessary from the point of view of a relation-based description of lexical meanings, but they appear helpful for human users and are used in several applications (starting with WSD) as well as in wordnet verification. The number of glosses was increased since 3.0 version by 6,445 and is 170 122, while the number of use examples was increased by 4,763 to 78 001. We assumed that not every LUs must be described by a use example, the priority is given to LUs of polysemous lemmas. However, we aim at achieving a state in which all LUs are characterised by stylistic registers (added to plWordNet at a later stage of its development, as initially it was meant to represent only general language).

Work on glosses and use examples meets the expectations of users who want plWordNet to be more similar to a traditional dictionary in terms of structure, but enriched with relational description. In addition, as already mentioned, the non-relational elements are also useful for natural language engineering.

3.3 Semi-automated Mapping onto Semantic Valence Lexicon

Walenty (Przepiórkowski et al., 2014) is a large lexico-semantic valence dictionary developed independently of plWordNet, but with a lot of cooperation between the two teams. This resulted in its schema referring to plWordNet LUs and semantic selectional preferences often annotated with plWordNet synsets (Hajnicz et al., 2016). Unfortunately, the old, 2.1 version of plWordNet was used for this purpose. Our goal was to automatically map the semantic roles of *Walenty* onto plWordNet in order to increase the density of its relations.

In contrast to FrameNet (Ruppenhofer et al., 2006), automatically linked to Princeton WordNet on the basis of similarity of paraphrases of its units and Princeton WordNet relations (Tonelli and Pighin, 2009), the linking between plWordNet and Walenty was done semi-automatically, with a lot of manual verification. First, we compared 2.1 and 3.0 versions of plWordNet and generated a list of plWordNet synsets whose content differed between the two versions. Next, two rounds of correction were carried out: automatic (based on the comparison of synset content and LU properties) and manual (for synsets which represented too big discrepancies between the two versions). In the latter case, we corrected the discrepancies. The differences between 2.1 and 4.0 synsets were mainly due to the introduction of new LUs or distinguishing new synsets as hyponyms or hypernyms of 2.1 synsets. The final mapping included 2,480 mappings from 2.1 to 4.0 synsets, which allowed us to introduce 17 new relation types to plWordNet. These relations are the equivalents of semantic roles described in the semantic layer of Walenty: *Theme*, *Condition*, *Path*, *Manner*, *Location*, *Purpose*, *Initiator*, *Recipient*, *Attribute*, *Instrument*, *Stimulus*, *Result*, *Measure*, *Time*, *Experiencer*, *Factor*, *Duration*. Both plWordNet and Walenty are the sources that are currently manually verified and corrected with respect to quality and completeness of entry description. The next stage of mapping between the resources was adding the relations on the basis of semantic description in Walenty and plWordNet, but only those with the "checked" status where there were no doubts about their quality and completeness description. In this way, plWordNet was enriched with 3,406 relation instances between plWordNet synsets, showing selectional preferences of units in the semantic layer of Walenty.

4 Alignment to English

A self-contained construction of plWordNet brought about the need of its later alignment to Princeton WordNet. The process started in 2012 and has been continued since then.

I-relation	V		N		Adv		Adj		Total	
I-relation	pl	en	pl	en	pl	en	pl	en	pl	en
I-synonymy	31955	1962	38699	38690	999	999	4338	4339	45991	45990
I-partial syn.	0	2	5821	5698	311	309	1493	1430	7625	7439
I-int.-reg. syn.	205	206	1847	1849	48	48	95	92	2195	2195
I-meronymy	0	1	10785	7944	0	0	0	0	10785	7945
I-hypernymy	79	3447	30736	82315	112	9897	375	44373	31302	140032
I-hyponymy	3433	79	82309	30740	9901	112	44389	374	140032	31305
I-holonymy	0	0	7945	10785	0	0	0	0	7945	10785
I-Type	0	0	7724	623	0	0	0	0	7724	623
I-Instance	0	0	623	7724	0	0	0	0	623	7724
I-allative	40	0	0	0	0	0	0	0	40	0
I-delimitive	157	0	0	0	0	0	0	0	157	0
I-excess	21	0	0	0	0	0	0	0	21	0
I-perdurative	12	0	0	0	0	0	0	0	12	0
I-anticausative	451	0	0	0	0	0	0	0	451	0
I-atenuative	102	0	0	0	0	0	0	0	102	0
I-cumulative	126	0	0	0	0	0	0	0	126	0
I-procesuality	6	0	0	0	0	0	0	0	6	0
I-completive	34	0	0	0	0	0	0	0	34	0
I-inchoative	64	0	0	0	0	0	0	0	64	0
I-distributive	313	0	0	0	0	0	0	0	313	0
I-iterative	37	0	0	0	0	0	0	0	37	0
I-terminative	6	0	0	0	0	0	0	0	6	0
I-ablative	18	0	0	0	0	0	0	0	18	0
I-causative	80	0	0	0	0	0	0	0	80	0
I-c-c-made-of	0	0	2	0	0	0	1067	0	1069	0
I-c-c-resembling	0	0	0	0	1	0	946	0	947	0
I-c-c-related-to	0	0	1	0	97	0	22697	0	22795	0
Total	7139	5697	186493	186368	11469	11365	75401	50608	280502	254038

Table 3: Interlingual relation counts

It took the form of manual mapping that is aligning wordnet nodes (synsets) corresponding in meanings and relation structures via a rich set of interlingual relations, (Rudnicka et al., 2012). It quickly turned out that interlingual synonymy (representing Simple Equivalence, cf. Vossen (2002)) is not enough to link two independently built resources for two quite different languages. English is an analytical Germanic language, while Polish a synthetic Slavic one. Therefore, other Complex Equivalence relations had to be resorted to. In Tab. 3, we present the full list of interlingual relations with their respective counts. The most frequent one is interlingual hyponymy and this tendency occurs across all parts of speech. In the latest 4.1 version of plWordNet, we have expanded the synset mapping between plWordNet and Princeton WordNet.

Moreover, we have also developed the methodology for a more fine-grained sense-level mapping and applied it to a substantial sample of noun lexical units. The methodology is based on a manual verification of the values of equivalence features. These include formal features such as number, countability and gender; semantic-pragmatic features such as sense, lexicalisation of concepts, register, collocations and co-text; and translational features such dictionary listing, dictionary equivalent position, and translation probability. The features are used to define three types of equivalence links: strong, regular and weak.

5 Applications

plWordNet is available on open licence and has been downloaded by more than 1,100 registered users (both individual and institutional). It has also had a quite large number of non-registered users and tens of thousands of users of the on-line browser[2]. On the basis of citations, questionnaires of the registered users, and direct co-operation with users within CLARIN, we can attempt an overview of plWordNet 4.1 applications. First, it was applied in linguistics for an analysis of lexico-semantic fields (Stanulewicz, 2010), analysis of word-forming nests (Lango et al., 2018), derivational processes (Kyjánek, 2018),

[2] `http://plwordnet.pwr.edu.pl`

identification of semantic classes (Lis, 2012), study on multi-word expressions (support for their extraction, recognition, classification) (Mykowiecka and Marciniak, 2012), and measuring semantic similarity of words (on the basis of their relation structure) (Siemiński, 2012). It found several applications in bilingual lexicography, e.g. in the study on partial equivalences in bilingual dictionaries (Liu, 2018), building a multilingual dictionary of the Yiddish language, as well as development of several bilingual and multilingual dictionaries (Sosnowski and Koseska-Toszewa, 2015). plWordNet was used in applied linguistics, e.g. in studies on the second language learning (Madej and Kiermasz, 2015), clinical research in the lexical system and its disfunction of patients suffering from dementia and Alzheimer disease. It was also utilised in Social Sciences, including an analysis of the language in Polish social media (digital trace, speaker intention, content of blog posts) (Haniewicz et al., 2014, Wawer and Sarzyńska, 2018), analysis of personal self-descriptions (structure and content), and analysis of commercials in media Iwińska-Knop and Krystyańczuk (2016). plWordNet was used to construct new resources, e.g. the system of Polish National Library descriptors was mapped on it, and KPWr Corpus (Broda et al., 2012), Składnica Corpus (Woliński et al., 2011) were annotated by the selected LUs. The most numerous group are applications in Natural Language Engineering, e.g. evaluation of word embedding models on the basis of synonymy tests automatically generated from plWordNet (Piasecki et al., 2018), named entity recognition, text mining and semantic search (Maciołek and Dobrowolski, 2013), text classification and text relation recognition (Brzeski and Boiński, 2014), semantic indexing of text (Karwowski et al., 2018), assignment of descriptive keywords to text documents (as knowledge basis and keyword repository) (Kaleta, 2014), automated structuring of text data (Maciołek and Dobrowolski, 2010), text interpretation in chat bots, text semantic similarity calculation (Siemiński, 2012), anti-plagiarism systems (Szmit, 2017), generation of semantically related families/sets of words for Information Retrieval and Internet monitoring and text normalisation, e.g. in the legal domain (Pełech-Pilichowski et al., 2014). As plWordNet is expanded with emotive annotation, cf (Zaśko-Zielińska et al., 2015), it has been applied several times in sentiment analysis and development of sentiment lexicons (Rybiński, 2017). Finally, it was used in *Jasnopis* system for the analysis of text difficulty to extract synonyms and hypernyms of words classified as too difficult for the intended text difficulty level (Dębowski et al., 2015).

6 Further Works

Is it ever possible to complete a wordnet? plWordNet 4.1 size and coverage, as well as its growth since 3.0 version may suggest that it is. However, this is misleading. In the case of a very large wordnet, the focus shifts from mere growth to the improvement of the amount and quality of information expressed for different lexical meanings. We plan to continue the work on increasing the density of relations (especially for LUs described so far by a few, if not single links), continuous maintenance of the wordnet quality by encompassing new lemmas and LUs in a corpus-based way. Instead of incorporating more and more specialist vocabulary, we plan to develop a system of cross-resource mappings envisaged in (Maziarz and Piasecki, 2018) in order to build a system of terminological, ontological and knowledge resources around plWordNet and make it an interface between them and the natural language lexicon. In addition, we also plan to further expand the relation structure towards better support for Word Sense Disambiguation. Moreover, we are going to continue the works on sense-level mappings. While proceeding with manual mapping, we are also going to develop a semi-automatic prompt system.

Acknowledgements

The work co-financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education and the project funded by the National Science Centre, Poland under the grant agreement No UMO-2015/18/M/HS2/00100.

References

Eneko Agirre and German Rigau. Word sense disambiguation using conceptual density. In *Proceedings* *of the 16th conference on Computational linguistics-Volume 1*, pages 16–22. Association for Computational Linguistics, 1996.

Bartosz Broda, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardyński. Kpwr: Towards a free corpus of polish. In *Proceedings of LREC*, volume 12, 2012.

Adam Brzeski and Tomasz Boiński. Towards facts extraction from texts in polish language. 2014.

Łukasz Dębowski, Bartosz Broda, Bartłomiej Nitoń, and Edyta Charzyńska. Jasnopis–a program to compute readability of texts in polish based on psycholinguistic research1. *Natural Language Processing and Cognitive Science* _, page 51, 2015.

Agnieszka Dziob and Maciej Piasecki. Implementation of the verb model in plWordNet 4.0. In *Proceedings of the 9th Global Wordnet Conference*, 2018a. URL `https://pdfs.semanticscholar.org/af21/13bb896f08993f995a68bbfa0ff805e1cbcd.pdf`.

Agnieszka Dziob and Maciej Piasecki. Dynamic verbs in the wordnet of polish. *Cognitive Studies / Études cognitives*, 18, 2018b. URL `https://ispan.waw.pl/journals/index.php/cs-ec/issue/view/98/showToc`.

Agnieszka Dziob, Maciej Piasecki, Marek Maziarz, Justyna Wieczorek, and Marta Dobrowolska-Pigoń. Towards revised system of verb wordnet relations for polish. In *Proceedings of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets*, 2017. URL `ceur-ws.org/Vol-1899/CfWNs_2017_proc6-paper_7.pdf`.

Elzbieta Hajnicz, Anna Andrzejczuk, and Tomasz Bartosiak. Semantic layer of the valence dictionary of polish walenty. In *LREC*, 2016.

Konstanty Haniewicz, Monika Kaczmarek, Magdalena Adamczyk, and Wojciech Rutkowski. A case study of sentiment orientation identification for polish texts. In *2014 European Network Intelligence Conference*, pages 46–51. IEEE, 2014.

Krystyna Iwińska-Knop and Hanna Krystyańczuk. Wykorzystanie big data w badaniu wizerunku marki w świadomości konsumentów. *Ekonomika i Organizacja Przedsiębiorstwa*, (9):28–42, 2016.

Fredrik Johansson, Joel Brynielsson, and Maribel Narganes Quijano. Estimating citizen alertness in crises using social media monitoring and analysis. In *2012 European Intelligence and Security Informatics Conference*, pages 189–196. IEEE, 2012.

Zbigniew Kaleta. Semantic text indexing. *Computer Science*, 15, 2014.

Waldemar Karwowski, Arkadiusz Orłowski, and Marian Rusek. Applications of multilingual thesauri for the texts indexing in the field of agriculture. In *International Multi-Conference on Advanced Computer Systems*, pages 185–195. Springer, 2018.

Paweł Kędzia, Maciej Piasecki, and Marlena Orlińska. Word sense disambiguation based on large scale polish clarin heterogeneous lexical resources. *Cognitive Studies| Études cognitives*, (15), 2015.

Lukáš Kyjánek. Morphological resources of derivational word-formation relations. Technical Report 61 (2018): 49, ÚFAL MFF, Charles Univesity, Prague, 2018.

Mateusz Lango, Magda Sevcikova, and Zdeněk Žabokrtskỳ. Semi-automatic construction of word-formation networks (for polish and spanish). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, 2018.

Roman Laskowski. Kategorie morfologiczne języka polskiego – charakterystyka funkcjonalna. In Renata Grzegorczykowa, Laskowski Roman, and Henryk Wróbel, editors, *Gramatyka współczesnego języka polskiego. Morfologia*, pages 151–224. Warszawa: Wydawnictwo Naukowe PWN, 1998.

Claudia Leacock, George A Miller, and Martin Chodorow. Using corpus statistics and wordnet relations for sense identification. *Computational Linguistics*, 24(1):147–165, 1998.

Magdalena Lis. Polish Multimodal Corpus - a collection of referential gestures. pages 1108–1113. European Language Resources Association, 2012. ISBN 978-2-9517408-7-7.

Lixiang Liu. Partial equivalences in bilingual dictionaries: Classification, causes and compensations. *Lingua*, 214:11–27, 2018.

Przemysław Maciołek and Grzegorz Dobrowolski. Is shallow semantic analysis really that shallow? a study on improving text classification performance. In *Proceedings of the International Multiconference on Computer Science and Information Technology*, pages 455–460. IEEE, 2010.

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

Monika Madej and Zuzanna Kiermasz. Exploring the attitudes towards word clouds in junior high students with a different multiple intelligence type: A research project. In M. Marczak and M. Hinton, editors, *Contemporary English Language Teaching and Research*, pages 139–157. Newcastle: Cambridge Scholars Publishing, 2015.

Marek Maziarz and Maciej Piasecki. Towards mapping thesauri onto plWordNet. In Francis Bond, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the 9th Global Wordnet Conference, Singapore, 8-12 January 2018*. Global WordNet Association, 2018.

Marek Maziarz, Maciej Piasecki, Joanna Rabiega-Wiśniewska, and Stanisław Szpakowicz. Semantic Relations among Nouns in Polish WordNet Grounded in Lexicographic and Semantic Tradition. *Cognitive Studies*, 11:161–181, 2011. http://www.eecs.uottawa.ca/ szpak/pub/Maziarz_et_al_CS2011a.pdf.

Marek Maziarz, Stanisław Szpakowicz, and Maciej Piasecki. Semantic relations among adjectives in polish wordnet 2.0: a new relation set, discussion and evaluation. *Cognitive Studies*, (12), 2012.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, and Stan Szpakowicz. 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 Recent Advances in Natural Language Processing RANLP 2013*, pages 443–452. INCOMA Ltd. Shoumen, BULGARIA, 2013a.

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

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Paweł Kędzia. plwordnet 3.0 – a comprehensive lexical-semantic resource. In Nicoletta Calzolari, Yuji Matsumoto, and Rashmi Prasad, editors, *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2259–2268. ACL, ACL, 2016. URL http://aclweb.org/anthology/C/C16/.

Diana McCarthy and John Carroll. Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences. *Computational Linguistics*, 29(4):639–654, 2003.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to WordNet: An On-Line Lexical Database. *Int. J. of Lexicography*, 3(4):235–244, 1990.

Agnieszka Mykowiecka and Małgorzata Marciniak. Combining Wordnet and Morphosyntactic Information in Terminology Clustering. In *Proc. COLING 2012: Technical Papers COLING 2012, Mumbai, December 2012.*, pages 1951–1962, 2012.

Tomasz Naskręt, Agnieszka Dziob, Maciej Piasecki, Chakaveh Saedi, and António Branco. Wordnet-Loom – a multilingual wordnet editing system focused on graph-based presentation. In Francis Bond, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the 9th Global Wordnet Conference, Singapore, 8-12 January 2018*. Global Wordnet Association, 2018.

Tomasz Pełech-Pilichowski, Wojciech Cyrul, and Piotr Potiopa. On problems of automatic legal texts processing and information acquiring from normative acts. In *Advances in business ICT*, pages 53–67. Springer, 2014.

Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław, 2009. URL http://www.dbc.wroc.pl/dlibra/docmetadata?id=4220&from=publication.

Maciej Piasecki, Łukasz Burdka, Marek Maziarz, and Michał Kaliński. Diagnostic tools in plwordnet development process. In Zygmunt Vetulani, Hans Uszkoreit, and Marek Kubis, editors, *Human Language Technology. Challenges for Computer Science and Linguistics*, volume 9561 of *LNCS*, pages 255–273. Springer, 2016. doi: 10.1007/978-3-319-43808-5_20.

Maciej Piasecki, Gabriela Czachor, Arkadiusz Janz, Dominik Kaszewski, and Paweł Kedzia. Wordnet-based evaluation of large distributional models for polish. In *Proceedings of the 9th Global Wordnet Conference (GWC 2018). Global WordNet Association*, 2018.

Adam Przepiórkowski, Elżbieta Hajnicz, Agnieszka Patejuk, Marcin Woliński, Filip Skwarski, and Marek Świdziński. Walenty: Towards a comprehensive valence dictionary of Polish. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 2785–2792. ELRA, 2014.

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

Josef Ruppenhofer, Michael Ellsworth, Miriam RL Petruck, Christopher R Johnson, and Jan Scheffczyk. Framenet ii: Extended theory and practice. 2006. URL https://framenet2.icsi.berkeley.edu/docs/r1.7/book.pdf.

Krzysztof Rybiński. Sentiment analysis of polish politicians. *e-Politikon. Kwartalnik Naukowy Ośrodka Analiz Politologicznych Uniwersytetu Warszawskiego*, XXIV:162–195, 2017.

Sam Scott and Stan Matwin. Text classification using wordnet hypernyms. *Usage of WordNet in Natural Language Processing Systems*, 1998.

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

Wojciech Paweł Sosnowski and Violetta Koseska-Toszewa. Multilingualism and dictionaries. *Cognitive Studies| Études cognitives*, (15), 2015.

Danuta Stanulewicz. Polish terms for 'blue' in the perspective of vantage theory. *Language Sciences*, 32 (2):184 – 195, 2010.

Radosław Szmit. Fast plagiarism detection in large-scale data. In Cham: Springer, editor, *International Conference: Beyond Databases, Architectures and Structures*, pages 329–343, 2017.

Sara Tonelli and Daniele Pighin. New features for framenet: Wordnet mapping. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 219–227. Association for Computational Linguistics, 2009.

Giannis Varelas, Epimenidis Voutsakis, Paraskevi Raftopoulou, Euripides GM Petrakis, and Evangelos E Milios. Semantic similarity methods in wordnet and their application to information retrieval on the web. In *Proceedings of the 7th annual ACM international workshop on Web information and data management*, pages 10–16. ACM, 2005.

Ellen M Voorhees. Using wordnet for text retrieval. *WordNet: an electronic lexical database*, pages 285–303, 1998.

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

Aleksander Wawer and Justyna Sarzyńska. Do we need word sense disambiguation for lcm tagging? In *International Conference on Text, Speech, and Dialogue*, pages 197–204. Springer, 2018.

Marcin Woliński, Katarzyna Głowińska, and Marek Świdziński. A preliminary version of składnica—a treebank of polish. In *Proceedings of the 5th Language & Technology Conference, Poznań*, pages 299–303, 2011.

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