plWordNet 5.0 – challenges of a life-long wordnet development process

Ewa Rudnicka, Bartłomiej Alberski, Maciej Piasecki
Department of Artificial Intelligence
Wrocław University of Science and Technology
ewa.rudnicka, bartlomiej.alberski, maciej.piasecki@pwr.edu

Abstract

The construction of plWordNet began in 2005 and has been continued since then. In this paper we present the latest 5.0 version and describe the challenges connected with a life-long wordnet development process. These involve changes in the procedures and lexicographers' teams, the necessity to extend the lexical description and the need to link to external resources (Princeton WordNet, sense-tagged corpora, valence dictionary). We describe different strategies and diagnostics implemented to improve the quality of the resource.

1 Introduction

The most burning question for wordnets today is whether we still need them in the large language models (LLMs) era? With the advent of the Chat GPT and other dialogue models, more and more language data is automatically generated recently. The quality of such data varies. This creates a danger of affecting the actual language use by its native speakers. Also, once it is fed to the models as training data we end in a kind of a vicious circle (Balloccu et al., 2024). Therefore, high-quality manually crafted and curated lexical resources remain a valuable source of data for different purposes of AI and NLP, both development and evaluation. In this paper we present the newest version of plWordNet, 5.0., as the core of a system of inter-linked language resources.

plWordNet is a large, manually constructed, corpus-based wordnet of Polish, but over the years it has become more than that. Now it is the heart of a complex system of language resources encompassing sense-annotated corpora (KPWr, KPWr 100, Sherlock Holmes corpus, Wiki-GLEX, EmoGLEX, KGR10, Składnica, the corpus of examples from the valence dictionary Walenty, and a knowledge graph VeSNet – a network of interlinked thesauri, encyclopedias, ontologies and dictionaries. Moreover, it is linked to the Polish valence dictionary Walenty. Sense-annotated corpora constitute crucial resources for word-sense disambiguation systems and, currently, LLMs. Knowledge graphs like VeSNet provide reliable information both about language including its specialist domain and the world.

The paper will focus on three main directions of the latest plWordNet development, all of which are intertwined: reviewing the granularity of plWordNet senses, acquiring new (glosses) and usage examples from the external inter-linked resources such as the sense-annotated corpus KPWr and the valence dictionary Walenty, and verifying and modifying plWordNet relational structure.

2 Background

Born in the 80-ties, no longer developed, with gaps and flaws, Princeton WordNet remains a basic reference to English sense inventory in the NLP (Miller et al., 1990) (Miller and Fellbaum, 2007). One of the reasons for this fact is that it is linked to hundreds, if not thousands of other resources in English and other languages. These include ontologies (e.g. SUMO), corpora (SemCor, WordNet Gloss Tag) (Vial et al., 2017)), valence dictionaries (VerbNet, FrameNet) (Baker et al., 1998) (Ryant and Kipper, 2004), and, above all, wordnets of other languages (EuroWordNet, OpenMultiLingual WordNet) (Vossen, 1998a)(Vossen, 1998b)(Bond and Paik, 2012). In this way, a multi-lingual, multi-dimensional system of interlinked language resources was created, with great potential of use in the NLP. To name the main use cases, sense-anotated corpora are explored in word sense disambiguation systems, knowledge graphs in named entity recognition tasks. All of the above resources can be used as reliable sources of data in the validation of the workings of large language models.

Sense annotated corpora are very special, valu-

able type of corpora, because they combine the data of the actual word use with the information of their dictionary meaning. Building a good-quality resource of this type requires time and means. The first corpus anotated with WordNet senses was SemCor (Miller et al., 1993).

plWordNet started off in 2005 (Piasecki, 2009), while its mapping to Princeton WordNet in 2012 (Rudnicka et al., 2012). It had a deep corpus connection from the very beginning, since the KGR corpus served as its crucial source of information on the frequency of words, their senses and relations between them. This was possible due to a custom-designed method of (semi)-automatic extraction of lexico-semantic information from a corpus developed by (Piasecki, 2009). One of the perks of the method was a number of corpus examples added to plWN senses. Nevertheless, this process was still lemma, not sense based leaving the user find the appropriate example for a given sense. At a later stage of work, corpus examples were manually disambiguated by linguists and added to the proper LUs. The first Polish corpus annotated with plWN sense was the KPWr corpus (The Corpus of Wrocław University of Science and Technology) (Broda et al., 2012).

3 Verification and LU description enrichment process

In this section we describe the key stages of the latest plWordNet development and improvement process involving the selection of lexical units for manual verification and description enrichment.

Originally, plWordNet was built of lexical units, synsets and relations between them. Additional elements of lexical description such as glosses, registers and usage examples were scarce. Examples and glosses were added only in the case of polysemous and hard to distinguish senses in order to clarify their meanings. However, at some point, purely relational description became insufficient for many, especially users, e.g. researchers from Social Sciences and Humanities (SS&H) who began to treat plWN as a kind of electronic dictionary. Furthermore, the presence of glosses and examples improves the quality of NLP (Bevilacqua et al., 2020)(Janz and Piasecki, 2023)(Banerjee et al., 2003). With the growing awareness of the role of glosses, registers and usage examples, they started to be systematically added to the description of lexical units (ver. 3.0, (Maziarz et al., 2016).

Lemmas	194 107
LUs	294 842
Glosses	160 937
Examples	240 741
LUs relations	275 367
Synsets	228 308
Synset relations	427 921

Table 1: plWordNet 4.2 statistics

Still, knowing the plWN size, it was clear that it was going to be a long-term task.

3.1 Selection of lexical units for verification

The starting point for our work was plWordNet 4.2 version. The essential statistics for this version are given in Table 1.

When we juxtapose the number of LUs (ca 290k) with the number of glosses (ca 170k glosses), we observe that only slightly above a half of LUs have glosses. The number of examples (ca 241k) is comparably higher, but it still does not cover all LUs. In addition, some LUs may have more than one usage example, while other ones none at all. Therefore, in our recent work we set off to fill in at least the part of the missing glosses and examples. To that end we decided to capitalise on the existing connections between plWordNet senses and different corpora.

Bearing in mind the size and complexity of plWordNet database as well as the time of its development, choosing the subset of units for verification and description enrichment has been a non-trivial task. In the latest stage of the plWordNet improvement process, we made a systematic selection of lexical units for verification and description enrichment. We decided to focus on the following four groups of lexical units:

1. LUs with a direct link to sense-tagged corpora, yet without glosses or examples (ca 13k);
2. LUs without glosses and examples and not attested in the sense-tagged corpora (ca 23k);
3. LUs with glosses, but without examples and not attested in the sense-tagged corpora (ca 46.5k)
4. verbal LUs with a direct link to the Walenty valence dictionary (ca 3.1k).

The detailed numbers are given in Table 2.

LUs in sense-tagged corpora without glosses and examples	13 233
LUs without glosses and examples and with no attestation in the corpora	22 724
LUs with glosses, but without examples	46 523
Verbal units linked to Walenty	3096

Table 2: Selected LUs

Alongside those four main groups, we have been working on the so called 'problematic cases' tagged as 'to be verified' in the WordNetLoom editing system. They originally came from our internal worksheets produced during other works, such as, for instance, corpus annotation with senses. The problems were related to the suspected wrong granularity of senses, wrongly directed relations, wrong usage example(s) or a mistake in the gloss (too wide or too narrow), wrong collocation provided for a given sense, or improper semantic domain ascribed. There were about 1k of such problematic LUs marked in the internal worksheets and about 6.5k marked directly in the database using the WordNet Loom 2.0 application. Still, it must be emphasized that the signal for verification did not necessarily mean that correction was needed in each case.

Having chosen a subset of lexical units for description enrichment and verification, we decided to work from the lemma level going two directions. On the one hand, we were checking the granularity of senses for each lemma. On the other hand, we were adding glosses and usage examples to individual lexical units which allowed us to verify the credibility of their sense granularity.

3.2 Reviewing the granularity of selected plWordNet senses

The primary source of mistakes in plWordNet is the wrong granularity of senses that is distinguishing too many or too few senses for a given lemma. The effect of such mistake is too narrow or too broad meaning of a given lexical unit which results in senses overlapping the scope of their meanings. Therefore, the verification of a potential mistake always needs to start from checking all the existing senses of a given lemma, especially with reference to the corpus data. The next step is to provide the gloss, register and at least two usage examples for a given sense. Lexical units with all those elements present will be classified as units with the higher description standard. In the course of work, we added almost 60k new usage examples and 8K new glosses as shown in Table 3.

Lexical units for whose senses no corpus attestation was found were marked for an extra verification at the later stages of plWN development. Currently, there are 1.2k of such units, that can be divided into the following groups:

1. imprisoned meanings, e.g. *etylowy*.1 'ethyl [alcohol]';
2. non-lexicalised multi-word expressions, e.g. *poczucie własnej godności*.1 'sense of self-dignity';
3. borrowings, e.g. *blezer*.1 'blazer';
4. contractions, *TB*.3 'terabite';
5. typos in lemmas, e.g. *leiszmania*.1 'Leishmania';
6. some participles, e.g. *dorastający*.1 'growing up';
7. nominalised adjectives, e.g. *małowierny*.1 'unfaithful';
8. gerunds, e.g. *obudzenie*.2 'waking up';
9. neologisms, e.g. *odsetka*.1 'percentage';
10. archaisms, e.g. *dębnik*.2 'tan'.

As for the verbal units, our works are carried out on the whole derivational nests, aspectual pairs and reflexive verbs with the reflexive particle *się*. We further process only such forms and senses that are attested in corpora. For example, the analysis of the verbal lemma *stawiać* will not only cover its 25 senses, but also its derivatives such as *postawić* 'put on', *ustawiać* 'put', *ustawić* 'put', *wystawać* 'stick out', *przedstawiać* 'present', *przedstawić* 'present', *przeciwstawiać* 'contrast', *przeciwstawić* 'contrast', *przeciwstawiać się* 'resist' and *przeciwstawić się* 'oppose'.

3.3 In search of new glosses and usage examples

The procedure of adding glosses and examples was fully manual. The senses selected for verification (see Table 2, Sect. 3.1) were divided according to their part of speech and the number of meanings of a lemma. After a given sense was identified and attested in the corpora, missing glosses and/or usage

examples were added. The work of the lexicographers was supervised by a coordinator. Verified usage examples were automatically added to the database (in three stages). Examples that were directly added to the database were verified and corrected in the database (in the WordNetLoom system). The most common mistakes included:

1. wrongly identified meaning, e.g. *we własnym sosie* 2 – oppressively 1, that is 'w zaduchu' 'in the chokehold', while the found usage example as its illustration corresponds to the different senses: *we własnym sosie* 1 – unalterably;

2. an improper member of the aspectual pair, e.g. *wtargać* 1 – bring 1, while the usage example includes the verb *wtargnąć*;

3. the absence of the actually described LU in the usage example, e.g. in the case of *rozbójnictwo* 1 – banditry 1, the provided examples include *rozbójnik* 1 – bandit 1 or *rozbój1* – mugging 1;

4. improper verbal form, e.g. *wykwitnąć* 3 – pop out 1 – an infinitive, while in the usage example a participial form is used "były wykwitającymi kwiatami" – 'were blooming flowers' or e.g. *zesłodować* 1 – malt 3 – an infinitive, while the usage example includes *zesłodowanie* — a gerund form;

5. metalanguage in usage examples, e.g. *świński trucht* 2, while the example goes:
   Pierwszy raz słyszę, by ktoś, narzekając na wolny przebieg spraw, mówił o "świńskim truchcie".
   'It is the first time that I hear somebody complaining about the slow pace of process and speaking about "(*approx.*) jog trot"'
   – in this particular case, the expression *świński trucht* is used in a metalanguage function, not literally as a part of the sentence. It is even emphasised by the use of quotation marks and the introduction "speaking about".

The results of the work are described in Table 3.

To conclude, we set the required standard of description at the level of lexical units so that each lexical unit is assigned its register, definition, and also minimum two use examples. All lexical units verified to meet this standard receive the status of partially processed. As a result of recent work, the number of units with the highest description standard has almost tripled, and almost 30k (precisely 29 367) lexical units have achieved it. This resulted in the manual addition of almost 60k (56 641) examples and almost 8k (7 561) definitions to the lexicon. Before this phase of work had been started, glosses sometimes were one-word only, and many lexical units had no definition at all. It was especially common practice in the case of monosemous lemmas.

LUs in sense-tagged corpora without glosses and examples	5768
LUs without glosses and examples and with no attestation in the corpora	19 523
LUs with glosses, but without examples	46 523
Verbal units linked to Walenty	864
New examples	56979
New glosses	7561
LUs in new standard	29518
LUs to be verified in the next stage of work	1274

Table 3: Results of the verification and enrichment process

4 Verification and improvement of the plWordNet sense and synset relational structure

4.1 Verifying the hypo and hypernymy relation structure

The backbone of the plWordNet (hierarchical) vertical structure is mainly formed by hyponymy and hypernymy relations. They form a bidirectional pair and are as the main *constitutive relations* (Piasecki, 2009)(Maziarz et al., 2013) for all parts of speech. In short, constitutive relations are a subset of lexico-semantic relations that determine by definition the wordnet structure and serve as a basis for defining synsets, i.e. lexical units sharing relation structures are grouped into synsets, see (Maziarz et al., 2013). Thus, analysis of the local structure of constitutive relations reveals if a given word sense, represented by a lexical unit, is correctly described. Further more, comparison of such local constitutive relation structures for the lexical units of a given lemma provides insight into proper identification of the different senses. Thus, in order to properly characterise a lexical unit it is necessary to recog-

nise and describe its relations with other lexical units that results in its inclusion into a synset (one lexical unit belongs to one synset only).

The set of required relations of plWordNet has been slightly evolving over years, and its contemporary state is presented below (Dziob et al., 2019):

1. for all parts of speech: hypo/hypernymy, mero/holonymy and inter-register synonymy;
2. in addition, for adjectives and adverbs: value of the attribute;
3. verbs, see (Dziob and Piasecki, 2018): presupposition, preceding, meronymy/holonymy, inchoativity, causality, pro- cessuality and state.;
4. for proper nouns only: type/instance.
5. relational adjectives are described only by relacyjność
6. feminine nouns are described by the femininity relation.

During the verification of the correctness of constitutive relations it often turns out that the hypernym for a given synset is semantically too wide, i.e. too high in the hypernymy structure (too close to top level), or too narrow (too deep in a subtree). A consequence of the incorrect hypernymy scope is wrong sense description of the hyponyms of a lexical unit. In the case of too high hypernymy location of a lexical unit too general meaning specification may be ascribed to its hyponyms, especially when their remaining relation structure is poor or even does not exist. Such a situation would be a hypernym: człowiek ze względu na swoje zajęcie1 for kaletnik 1 – skinner 4. A much better hypernym for kaletnik 1 is rzemieślnik 1 – craftsman.3. Careful inspection of the hypo/hypernymy structure resulted in 115 473 changes introduced in plWordNet 5.0 in comparison to the previous version.

4.2 Increasing the relation structure density

In addition to the verification of the existing relation structure, another important direction in improving the wordnet quality is increasing the relation structure density – the structure is the primary means for expressing knowledge about word senses. In the earlier versions of plWordNet, at least one constitutive relation was considered satisfactory for a minimal description of an lexical unit. However, recently we aim at increasing the number of relations per a single LU. Literature studies show that the quality of text processing increases with the increase in the number of relations per lexical unit in a lexical resource used for the purposes of processing (Bevilacqua et al., 2020) (Janz and Piasecki, 2023). Therefore, during the verification of selected nodes of the wordnet graph we add new relations from a wide spectrum of relations available in plWordNet. These involve both synset and sense-level relations. The results of our work are shown in Table 4 where we juxtapose relation counts for 4.2 and 5.0 versions of plWordNet. The number of sense relations have grown by 3.5k, while the number of synset relations by 43k.

Number of relations	plWN 4.2	plWN 5.0
LUs	275 367	278 934
Synsets	567 871	610 806

Table 4: plWordNet 4.2 and 5.0 sense and synset relation counts

5 Verification and improvement of interlingual links

The manual mapping of plWordNet onto Princeton WordNet has been carried out since 2012 (Rudnicka et al., 2012, 2021). It was a dynamic process with mapping procedures refined or modified in response to results of the earlier mapping or new mapping challenges. At the beginning, lexicographers had to mainly rely on the internal relation structures of plWordNet and Princeton WordNet as well as glosses and usage examples if such were available. As the network of interlingual relations was growing they gained additional information – the existing interlingual relations whose input also needed to be taken into account while establishing new relations. Throughout this time, there has been also many changes in plWordNet itself. Thus, there are certainly nodes or fragments of the bilingual wordnet graph which could be improved.

Therefore, to address these issues, we have designed a series of automatic diagnostics of the interlingual relation system that were run through the database. Next, their results in the form of the produced lists of synsets and lexical units were presented to lexicographers in the Tracker system (Naskręt et al., 2018). They manually analysed them and, in consequence, some links were deleted, other ones altered.

5.1 Synsets

The first series of diagnostics was designed to eliminate the obvious mistakes, the kind of 'slips of the tongue' or 'typos', which bearing in mind the time and scope of manual work were bound to happen occasionally. Those involved the following:

1. links between synsets of improper parts of speech (e.g. I-mero/I-holonymy between non-noun synsets);
2. wrong direction of a relation (PL-ENG/ENG-PL);
3. missing bidirectional relations (e.g. I-hypo/I-hypernymy).

Since the beginning of synset mapping, we have followed (Vossen, 1998b) in assuming one interlingual synonymy relation per synset. This restriction was lifted for verbal synsets due to essential differences in lexicalising aspect in English and Polish (lexical aspect in Polish vs grammmatical aspect in English). Consequently, the pairs of Polish perfective and imperfective verbal synsets were allowed to be mapped to the same English verbal synset (covering perfective and imperfective senses of a given verb) (Rudnicka et al., 2021). Still, our first diagnostic test was to check if there were any instances of multiple synonymy links between Polish and English synsets and verify them manually. We deleted 1 167 mistakes and left 2 712 links between verbal synsets that were correct. 1 079 new links were added. The number of deleted links is very small in comparison to the overall number of I-synonymy links which amounts to 943k.

Another diagnostic directly linked to the mapping procedure was checking the cases of a simultaneous interlingual synonymy and interlingual hyponymy links for a single synset. Interlingual synonymy was always treated as a priority relation in the mapping procedure. If it could be established no further relation was necessary. Interlingual hyponymy was only introduced when no interlingual synonym could be found in the other wordnet. However, since the mapping was extended in time and carried out by partially different lexicographers' teams at different stages there could appear situations when both relations existed for the same synset linking it to two different other language synsets. For example, the Polish synset *farsowość*.1 was linked to the English synset *comicality*.1 via interlingual hyponymy relation and to *farcicality*.1 via interlingual synonymy relation. Since

Interlingual relations	deleted	added
verbal synsets	1 065	19 816
non-verbal synsets	14 868	23 972
Sum	15 933	43 788

Table 5: interlingual synset relations for verbs and other POS

I-synonymy is more detailed than the I-hyponymy we deleted the latter. Nevertheless, there may be cases where such pairs of relations are justified. This happens when the hyponymy relation is the only interlingual relation describing its hypernym. For example the English synset *interactive kiosk*.1 is the I-synonym of *infokiosk*.1, but also I-hyponym for *kiosk*.4, which is the only interlingual relation of the latter. All in all, as a result of verifying this diagnostic we deleted 1 737 relations and added 735 new ones.

Certain changes were connected with introducing new interlingual relations at further stages of work, such as, for instance interlingual inter-register synonymy for synsets sharing the meaning but differing in register. This relation started to be systemically introduced for Polish synsets which were stylistically marked and thus inter-register synonyms to neutral Polish synsets otherwise linked via I-synonymy to neutral English synsets. Before introducing this relation I-hyponymy was used, but it was later replaced with I-inter-register synonymy. We replaced 1 168 instances of such relation, i.e. to improve consistency of application of different relations.

Apart from analysing the results of our diagnostic tests, we have been also continuing the works on filling in the missing links between plWordNet and Princeton WordNet synsets. These mainly focused on verbal synsets and on Princeton WordNet synsets. Verbs were the last of all parts of speech that we started to map, while for a very long time we were going from plWordNet to Princeton WordNet direction which left a number of English synsets unmapped. The summary results of our work are shown in Table 5.

5.2 Lexical units

Equivalence relations between lexical units form an extra layer of interlingual mapping between plWordNet and Princeton WordNet (Rudnicka et al., 2019). They link pairs of Polish and English lexical units that display strong equivalence in meaning and use and thus function as (mutual)

domains	LU pairs	Right links	Wrong links	Deleted links	Changed links
location	35	31	0	1	3
activity	93	64	7	19	1
property	96	78	18	6	12
artefact	156	145	0	8	0
natural object	94	86	8	3	7
body part	93	86	0	6	3
thinking	81	68	13	0	13
group	37	29	0	6	4
plant	31	28	3	0	3
state	53	46	0	4	3
communication	64	51	0	11	2
relation	41	32	0	9	1
food	30	23	7	1	6
natural phenomenon	39	34	5	0	6
possession	54	46	0	8	0
event	38	33	5	0	5
natural proc.	43	33	2	8	0
quantity	43	39	0	3	0
substance	31	29	1	1	0
emotion	24	10	1	1	1
system	13	11	0	1	0
animal	8	7	0	1	0
time	8	8	0	0	0
shape	13	12	1	0	1
person	6	2	1	0	2
h. hierarchy	3	3	0	0	0
purpose	4	2	1	0	0
SUM	1231	978	73	97	73

Table 6: Manual verification of equivalence mapping across domains

translational equivalents. This applies especially to strong and regular equivalence links. In addition, weak equivalents links hold between units that can function as translational equivalents, even descriptive ones (Rudnicka et al., 2019).

The number of equivalence links is much smaller than that of interlingual relations for two reasons. First, it was not the goal to introduce such links for all lexical units, because this would not be possible due to substantial differences between languages (Polish and English). Second, the mapping procedure was even more demanding than that for the interlingual mapping between synsets, hence very time consuming. Currently, the equivalence mapping exists for almost 27k pairs of Polish and English noun lexical units.

In establishing so strong type of interlingual links as equivalence links one would expect the correspondence in semantic domains of the Polish and English lexical units. However, lexical unit assignment to both Princeton WordNet lexicographer's files and plWordNet domains is to some extent arbitrary and cannot be treated as a decisive factor in establishing a link. Still, some domains tend to correlate, others do not (Maziarz et al., 2014). Therefore, we have checked the distribution of domains between Polish and English lexical units linked by equivalence links and selected for manual verification such pairs of domains that occur five or less times in the mapping.

	deleted	added	sum
Equivalence relations	810	13 623	14 433

Table 7: Equivalence relations

The results of the manual verification of equivalence links across plWordNet domains are presented in Table 6. We observe that most of the analysed connections were right links. The exact shares vary across domains. This finding corroborates the prediction that domain assignment is arbitrary, especially across two wordnets of two very different languages. It can also be treated as a kind positive validation/re-evaluation of the equivalence mapping procedure used in linking plWordNet and Princeton WordNet senses. On the other hand, for each domain we have discovered some number of links that had to be altered, either deleted altogether or changed to a different type of equivalence link. In some cases the change also involved the change of the interlingual relation between the plWordNet and Princeton WordNet synsets the lexical units in question were the components of.

In addition to the verification of the earlier existing links, we have also continued with introducing new equivalence links. The summary results of our work are shown in Table 7.

6 Conclusion and future work

We plan to further increase the quality of plWordNet and provide each lexical unit with a higher standard of description that is a gloss and minimum two usage examples.

Other planned work is to continue work on the density of the relation network by increasing the number of instances and supplementing the list of relations with new types, e.g., the masculinity relation describing at the unit level masculine derivations derived from feminine word-forming bases, e.g., *wdowiec* 'widower' ⇒ *wdowa* 'widow', zodiakarz 'zodiacarius' ⇒ *zodiakara* 'zodiacara', or the compression relation linking at the unit level univerbisms with their word bases, e.g., *starówka* 'old town' ⇒ *stare miasto* 'old town'. It is also important to supplement the resource with new lexical units, as well as to verify and possibly correct selected parts of the graph.

It is also planned to integrate with the parallel Polish-English corpus Paralela (Pęzik, 2016), as well as with the Wielki Otwarty Korpus (WOK, Large Open Corpus of Polish) (Broda et al., 2012). The idea is that linking plWordNet to these resources will result in a very large sense-tagged corpus (for the purposes of word sense disambiguation). Another task is to create domain subwordnets, e.g., a subwordnet of musicological terms, which will be a separate domain resource, but will also be linked to plWordNet. First we will start with an experimental task that will test the theoretical assumptions and technical capabilities of such a solution. Ultimately, it will facilitate the NLP of domain texts.

Due to the technical possibilities of the WordNetLoom 2.0 editor (Naskręt et al., 2018) we plan to build test resources in a form of sub-wordnet. The current form of the tool makes it possible not only to add headwords that form a separate resource from plWordNet, but even to create your own types of relations or edit the existing ones. The planned sub-wordnets will be test wordnets including specialist vocabulary. The resource will be distinct from plWordNet, but connected to it via selected nodes consisting of common synsets. The enterprise has an experimental character.

Acknowledgments

References

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms. *arXiv preprint arXiv:2402.03927*.

Satanjeev Banerjee, Ted Pedersen, et al. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Ijcai*, volume 3, pages 805–810.

Michele Bevilacqua, Roberto Navigli, et al. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the conference-Association for Computational Linguistics. Meeting*, pages 2854–2864. Association for Computational Linguistics.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *proceedings of the 6th global WordNet conference (GWC 2012)*, pages 64–71. Matsue.

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

Agnieszka Dziob and Maciej Piasecki. 2018. Implementation of the verb model in plWordNet 4.0. In *Proceedings of the 9th Global Wordnet Conference*, pages 113–122, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.

Agnieszka Dziob, Maciej Piasecki, and Ewa Rudnicka. 2019. plWordNet 4.1 - a linguistically motivated, corpus-based bilingual resource. In *Proceedings of the 10th Global Wordnet Conference*, pages 353–362, Wroclaw, Poland. Global Wordnet Association.

Arkadiusz Janz and Maciej Piasecki. 2023. Word sense disambiguation based on iterative activation spreading with contextual embeddings for sense matching. In *Proceedings of the 12th Global Wordnet Conference*, pages 140–149.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, and Stan Szpakowicz. 2014. Registers in the system of semantic relations in plWordNet. In *Proceedings of the Seventh Global Wordnet Conference*, pages 330–337, Tartu, Estonia. University of Tartu Press.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Paweł Kędzia. 2016. plWordNet 3.0 – a comprehensive lexical-semantic resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2259–2268, Osaka, Japan. The COLING 2016 Organizing Committee.

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.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

George A Miller and Christiane Fellbaum. 2007. Wordnet then and now. *Language Resources and Evaluation*, 41:209–214.

George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Tomasz Naskręt, Agnieszka Dziob, Maciej Piasecki, Chakaveh Saedi, and António Branco. 2018. WordnetLoom – a multilingual Wordnet editing system focused on graph-based presentation. In *Proceedings of the 9th Global Wordnet Conference*, pages 190–199, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.

Piotr Pęzik. 2016. Paralela corpus and search engine.

M Piasecki. 2009. A wordnet from the ground up. *Oficyna Wydawnicza Politechniki Wroclawskiej*.

Ewa Rudnicka, Marek Maziarz, Maciej Piasecki, and Stan Szpakowicz. 2012. A strategy of mapping Polish WordNet onto Princeton WordNet. In *Proceedings of COLING 2012: Posters*, pages 1039–1048, Mumbai, India. The COLING 2012 Organizing Committee.

Ewa Rudnicka, Maciej Piasecki, Francis Bond, Łukasz Grabowski, and Tadeusz Piotrowski. 2019. Sense equivalence in plwordnet to princeton wordnet mapping. *International Journal of Lexicography*, 32(3):296–325.

Ewa Rudnicka, Wojciech Witkowski, and Maciej Piasecki. 2021. A (non)-perfect match: Mapping plWordNet onto PrincetonWordNet. In *Proceedings of the 11th Global Wordnet Conference*, pages 137–146, University of South Africa (UNISA). Global Wordnet Association.

Neville Ryant and Karin Kipper. 2004. Assigning xtag trees to verbnet. In *Proceedings of the 7th International Workshop on Tree Adjoining Grammar and Related Formalisms*, pages 194–198.

Loïc Vial, Benjamin Lecouteux, and Didier Schwab. 2017. *UFSAC: Unification of Sense Annotated Corpora and Tools*. Ph.D. thesis, UGA-Université Grenoble Alpes.

Piek Vossen. 1998a. Introduction to eurowordnet. *EuroWordNet: A multilingual database with lexical semantic networks*, pages 1–17.

Piek Vossen. 1998b. A multilingual database with lexical semantic networks. *Dordrecht: Kluwer Academic Publishers. doi*, 10:978–94.