plWordNet as the Cornerstone of a Toolkit of Lexico-semantic Resources

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

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

Abstract

A wordnet is many things to many people: a graph of inter-related lexicalised concepts, a taxonomy, a thesaurus, and so on. A wordnet makes good sense as the mainstay of any deep automated semantic analysis of text. We have begun the construction of a multi-component, multi-use toolkit of natural language processing tools with plWordNet, a very large Polish wordnet, at its centre. The components will include plWordNet and its mapping onto an ontology (the upper level and elements of the middle level), a lexicon of proper names and a semantic valency lexicon. Some of those elements will be aligned with plWordNet, and there will be a mapping onto Princeton WordNet. Several challenging applications will show the utility of the toolkit in practice.

1 How wordnets evolve

Wordnets start small but quickly grow to account for much of the lexical material of the given language. The size of version 3.1 of Princeton WordNet (PWN) (Fellbaum, 1998) is a *de facto* standard, even if this mature wordnet also keeps growing, albeit slowly.[1] One of the resources which approach this size standard is plWordNet (Piasecki et al., 2009), now in version 2.1. Languages change continually, so lexicographers never rest, but one can still ask when the development of a wordnet ought to slow down, and whether there is an appropriate steady state of a wordnet. That clearly is a loaded question, and much depends on the language. For example, suppose that a wordnet for a richly inflected language with complex and varied derivation was originally a translation of PWN. Such a wordnet should, sooner or later, acquire semantic relations which account accurately for its unique lexical system..

A wordnet, even as developed as PWN, GermaNet (Hamp and Feldweg, 1997) or plWordNet (Maziarz et al., 2013a), serves many natural language processing (NLP) applications, yet it seems neither feasible nor necessary to remake wordnets into universal NLP resources. Instead, we propose to mark clear boundaries around a wordnet (what it should and what it should not include), and treat it as a pivotal element of an organic toolkit of inter-connected tools and resources for the semantic analysis of texts, along with the auxiliary morphological and syntactic analysis tools. Our case study is such a toolkit, now under development, centred on plWordNet 3.0 (also in development), and intended first and foremost for research in the humanities.

In the remainder of the paper, we present the main design assumptions and principles of that project. We explain how comprehensive we want plWordNet 3.0 to become, what size and what coverage we envisage. We attempt to describe how the toolkit will be built around plWordNet, and we outline plans for its large-scale illustrative applications in several domains. We discuss how the components of the toolkit will be expanded or constructed: plWordNet 3.0, its mapping to an ontology, and a semantic lexicon of proper names. We also briefly present resources for morphological and structural description, as-

[1] PWN began as a test of a theory of human semantic representation and memory (Collins and Quillian, 1969). It now features a comprehensive vocabulary, a set of universally useful semantic relations, glosses, links to ontologies, and more.

sociated with the plWordNet system, among them a lexicon of lexico-syntactic structures of multiword expressions and a valency lexicon linked to plWordNet but developed independently.

This work is meant to take several years of initial effort and years of maintenance. We cannot answer many design questions *yet*, but many will be answered as the project unfolds. That is to say. we want to interlace theory and practice.

2 The cornerstone

2.1 The model of plWordNet

There is a rather unfortunate tendency to treat wordnets as a substitute for ontologies (which are perhaps less well known and less easily available to the NLP community), but significant differences are clear when one compares an ontology with a wordnet understood as a lexico-semantic resource (Prévot et al., 2010). A systems of concepts in a wordnet must be expressed entirely in a natural language – unlike ontologies. A strict knowledge representation is required in an ontology, but a wordnet works through words. The inherent ambiguity of the lexical material makes very formal definitions infeasible. In particular, synonymy is a matter of degree, while concepts in an ontology should be defined with certainty. A rigorous construction of an ontology is not easy insofar as language intuitions "get in the way". For example, PWN contains a network of conceptual relations between synsets which represent *lexicalised concepts*, but – unsurprisingly – no formal definition of the notion of *concept* has been put forward yet. PWN's structure was shaped by the lexico-semantic dependencies among words, not by formal properties of an ontology structure.[2]

Corpus analysis can help recognise lexico-semantic relations for inclusion in a wordnet. Practical substitution tests can be formulated for individual relations without committing to any particular theory of lexical semantic or human cognition, in the spirit of *minimal commitment* (Maziarz et al., 2013b). A wordnet so conceived provides a description of the lexical system which is well defined and grounded in language data. It can also be built up at a considerably low cost and with a high degree of consistency.

Corpus-based wordnet development, which has led to plWordNet 2.1, assumes a very large monolingual corpus as the main source of lexical knowledge. Software tools facilitate corpus browsing and semi-automatic knowledge extraction (Piasecki et al., 2009). Dictionaries and encyclopedias are consulted in order, if necessary. This rigorous procedure limits the variability of editing decisions by circumscribing the role of linguistic intuition, though intuition still has its place as a final recourse.

A wordnet based very closely on language data is easier to develop when its primitive is a *linguistically* motivated construct: the lemma-sense pair which we call the *lexical unit* (LU). The plWordNet model, described in detail in (Maziarz et al., 2013b), considers lexico-semantic relations between LUs. LUs are grouped into synsets if they share lexico-semantic relations from a pre-defined repertory, called *constitutive relations*. They must be fairly *frequent* (to describe many LUs), *shared* among LUs (to define groups), *grounded* in the linguistic tradition (to facilitate their consistent understanding) and, if possible, already *used* in other wordnets (to improve compatibility). One of the effects is that synonymy is not a primary relation. It is derived from other lexico-semantic relations, notably hyponymy and hypernymy, which are much simpler to recognise consistently. A relation between two synsets is directly derived from lexico-semantic relations, and it is effectively an abbreviation for a set of links defined for all pairs of LUs from both synsets.

Not every lexico-semantic relation qualifies as a constitutive relation. For example, antonymy is not shared widely enough, and there are no "co-antonyms" for the same LU. Antonymy obviously belongs in a wordnet, but not as a defining factor. Another example: plWordNet does not directly include derivational relations which describe transformations of the basic morphological word forms. It only records lexico-semantic relations signalled by those formal transformations. For example, the same morpheme can be used to create forms of different meanings, so in each case we describe a different specific lexico-semantic relation rather than the formal dependencies among word forms (Piasecki et al., 2012b).

When we wrote precise definitions and substitution tests, we realised that several factors systematically constrain linking large sub-classes of LUs by lexico-semantic relations. Three of those fac-

[2] Put another way, there can be a disconnect between the "straitjacket" of an ontology and the inevitable vagueness and context-dependence of actual texts.

tors, stylistic registers, verb aspect and semantic verb classes, apply frequently enough to allow explicit treatment in the relation definitions (Maziarz et al., 2013b). They refer to the properties of LUs, so we call them *constitutive features*. Relations strictly limited to verbs of the same aspect and semantic class include hyponymy and several specific entailment relations such as inchoativity. Registers explain many situations when pragmatic limitations prevent LUs with the same denotation from being used in the same contexts. Such LUs do share some relations, so constraining relation definitions by register compatibility helps shape the wordnet structure consistently.

Glosses may play a secondary role in a representation of lexical meaning based on the relational paradigm, but writing them helps wordnet editors work with polysemous lemmas. They are also helpful for human users and very useful in applications. Automatically extracted usage examples, equally secondary, are very popular with users in linguistics. We will, therefore, place plWordNet 3.0 glosses and examples in for as many LUs as possible, though the final numbers are hard to put now on this laborious process.

The system of lexico-semantic relations in plWordNet 3.0 will not differ much from plWordNet 2.1. The verb hypernymy structure putting verbs into semantic classes may have to be adjusted. The adverb network must be built from scratch. It will also be important to increase network density for the existing relation types.[3]

The whole plWordNet 3.0, together with all associated resources and mappings, will be naturally available on an *open* WordNet-style licence.

2.2 Size matters

Table 1 shows that plWordNet 2.1 comes close in size to PWN 3.1: nearly the same number of synsets, and about 2/3 of the lemmas and LUs. We want the vocabulary to correspond to the contents of a large morpho-syntactic dictionary (Saloni et al., 2012) commonly used when processing Polish texts, but the coverage is still far from that number.[4] The target size of plWordNet 3.0 is not easy to set *a priori*, but we know that it is better to count lemmas than synsets (assuming that all senses of a lemma are accounted for).[5] Note that infrequent words need a representation in wordnets more than frequent words, well described by knowledge automatically extracted from a large corpus. Measures of semantic relatedness tend to be useless for lemmas appearing less than 50 times in a corpus of more than 1 billion tokens (Piasecki et al., 2009). That said, it is unrealistic to aim for a wordnet with full coverage of a frequency list based on a very large corpus.

POS	synsets	lemmas	LUs	avs
N-PWN	82,115	117,798	146,347	1.78
N-plWN	80,950	78,184	110,913	1.37
V-PWN	13,767	11,529	25,047	1.81
V-plWN	21,770	17,518	32,037	1.47
A-PWN	18,156	21,785	30,004	1.65
A-plWN	15,113	11,651	18,748	1.25

Table 1: The count of Noun/Verb/Adjective synsets, lemmas and LUs by part of speech (POS), and average synset size (avs), in PWN 3.1 (PWN) and plWordNet 2.1 (plWN).

It is hard to say just how many words there are in a language, never mind newest coinage. Corpora, even huge, are not complete enough (Kornai, 2002; Gale and Sampson, 1995, p. 218). One might assess a lower bound of the vocabulary size from existing dictionary sizes, or calculate it analytically with corpus and statistical methods.

English is often assumed to have the most words. The Oxford English Dictionary (Simpson, 2013) contains 300k main entries (± lemmas) and 600k word forms, but no freshest neologisms. There are even larger dictionaries: *Woordenboek der Nederlandsche Taal* with 430k entries (Nijhoff, 2001) and a 330k dictionary of Grimm brothers (Grimm, 1999); both are contemporary *and* historical. A comparable Polish dictionary from the early 1900s has 280k entries (Karłowicz et al., 1900–1927; Piotrowski, 2003, p. 604). Modern dictionaries of general Polish have fewer entries: 130k (Zgółkowa, 1994–2005), 125k (180k LUs) (Doroszewski, 1963–1969), 100k (150k LUs) (Dubisz, 2004), 45k (100k LUs) (Bańko, 2000). They do not contain many specialised words and senses from science, technology, culture and so on, appropriate for a wordnet.

[3] There are 3.99 relations per noun synset, 3.06 relations per verb synset, 1.56 per adjective synset inplWordNet 2.1. In PWN: 3.54 for nouns, 2.21 for verbs and 2.43 for adjectives.

[4] (Saloni et al., 2012) has around 200,000 lexemes (our lemmas), but that includes many proper names.

[5] The number of lemmas covered tells how many out-of-vocabulary words to expect during processing.

corpus	corpus size	# entries
Cobuild (1986)	18M	19.8k
Cobuild Bank of English (1993)	121M	45.2k
Bank of English (2001)	450M	93.0k
plWordNet	1,800M	≈174.0k

Table 2: Dictionary size in entries as a function of corpus size according to Krishnamurthy. For comparison – the estimates for plWordNet.

Krishnamurthy (2002) ties the corpus size to the number of lemmas which occur 10+ times. We added an extrapolation for plWordNet (Table 2): 174k lemmas, a little more than we propose to have in plWordNet 3.0.[6]

If we could double our current corpus, the approximation in (Good and Toulmin, 1956; Efron and Thisted, 1975, eq. 2.7) would be useful:

$$\hat{\Delta} = \sum_{x=1}^{\infty} (-1)^{x+1} n_x,$$

$\hat{\Delta}$ is the size of a new vocabulary found in the new part of the corpus, n_x is number of word types used x times in the source corpus (before doubling). This gives 1,322,850 new word types for the doubled plWordNet corpus. Standard deviation is given by formula (2.10) in (Efron and Thisted, 1975) :

$$S = \sqrt{var\hat{\Delta}} = \sqrt{\sum_{x=1}^{\infty} n_x} \approx \pm 42\text{k word types.}$$

This approximation, however, takes into account proper names, foreign words, typos and so on (Kornai, 2002, p. 83), undesirable in our wordnet. Even if we conservatively assume 15% "real" words,[7] we can count on some 200k additional lemmas. Multi-word lexical units would not be included in that estimate. See Table 3 for details.

In the end, we set the target size of plWordNet 3.0 arbitrarily at 200,000 lemmas: a lot, but it accords with the largest Polish dictionaries and with corpus statistics – and with the policy of accounting for rare lemmas. The completion is expected at the end of 2015. The number of synsets (218,000) and LUs (250,000) has been estimated by extrapolating the lemma-LU-synset ratios in plWordNet 2.1.

	# entries
Polish dictionaries	100-250k
plWordNet corpus, 10+ lemmas [K]	174k
doubled plWordNet corpus, 0+ lemmas [GT]	+200k

Table 3: Potential lemma count for plWordNet. Estimates due to Krishnamurthy [K] and Good & Toulmin [GT].

The size of plWordNet has already far exceeded the vocabulary of the average Polish user – by design. A wordnet should outstrip traditional dictionaries if it is to be part of language tools which work on the Internet scale (with practically limitless vocabulary) and without the benefit of human language intuition. plWordNet 3.0 will be part of the CLARIN language technology infrastructure[8] aimed at delivering research tools for processing text and speech resources in the very broad domain of the humanities and social sciences.

Not all applications benefit from a large wordnet. Word-sense disambiguation may suffer if there are too many too fine sense distinctions, but the granularity of the senses and the size in lemmas are not strictly correlated. The former is more a matter of a construction decision, with relatively infrequent cases of a lemma of the general register assigned new specific senses.[9]

Wordnet construction based on knowledge extracted from a large corpus (Piasecki et al., 2009; Piasecki et al., 2012a) reaches its limits when the most frequent vocabulary has been accounted for.[10] A Polish corpus of significantly more than the present 1.8 billion words is much harder to make than it would be for English if one wants to preserve quality.[11] Pattern-based relation extraction, better with low frequencies, tend to be less complete and less productive than statistical distribution-based methods. We will have to supplement corpus data with knowledge from such structured text resources as Wikipedia.

[6] This estimation was given by a regression curve:

$$N_{10+} = 6.67t^{0.477} \approx 6.67\sqrt{t},$$

where t is the corpus size and N_{10+} is the number of words with 10 or more corpus occurrences; the coefficient of determination equals 0.996. The equality is of a power-law kind, as Guiraud's law (Guiraud, 1954).

[7] Indeed, we found 15 common words in a 100-word sample taken from the plWordNet corpus frequency list.

[8] See `http://nlp.pwr.wroc.pl/clarin` and `http://clarin.eu`

[9] A small example: *dryl* 'drill' means an exercise or an ape, the latter very rare.

[10] Any measure of semantic relatedness works fine for 1,000 occurrences per one billion words, deteriorates for 100 occurrences and practically fails for 10.

[11] Language errors and irregularities quickly decrease the quality of morpho-syntactic preprocessing.

2.3 The quality

The current phase of our long-term project begins with plWordNet 2.1: version 2.0 with improvements due to the application of automated diagnostic tools, and a continually growing mapping to PWN 3.1. The development of plWordNet has been consistently carried out in WordnetLoom, a wordnet editor with advanced graphical editing capabilities and a palette of corpus search, dictionary search, structure checking and bookkeeping tools (Piasecki et al., 2013). WordnetLoom imposes many constraints on the wordnet relation structures, but we have discovered that more is required. New rules include the following:

- simple structural errors, such as the presence of lexical units (LUs) without synsets or links without the obligatory inverse counterpart for symmetric relations;
- general semantic errors such as hypernymy and meronymy cycles, more than one relation linking a pair of synsets, or direct and indirect relations linking mutually a pair of synsets;
- specific semantic rules developed for selected domains and hypernymy branches.

3 The toolkit of lexico-semantic resources

3.1 Multi-word expressions

Multi-word Expressions (MWEs), a substantial part of the lexicon, are under-represented in dictionaries and on frequency list. With effective MWE detection, a very large corpus is the most reliable source of MWEs, but (inconveniently) morphological analysis handles their elements separately. We will expand the dictionary of lexico-morpho-syntactic MWE structures from (Kurc et al., 2012) to more than 60000 MWEs in a separate resource linked to plWordNet 3.0.

3.2 Proper names

We treat proper names (PNs) as separate from the lexicon: very few PNs are present in general dictionaries. That is why they do not belong in lexico-semantic resources. In particular, hyponymy does not really apply. An entity denoted by a PN is an *instance* of a *type*. PNs are primarily characterised by their referents, not by their semantic properties revealed in use examples. One must know the referent of the given PN in order to to interpret it unambiguously. The instance/type relations are not lexico-semantic relations, so PNs can in principle be linked directly to an ontology, not to a wordnet. There are, however, two arguments in favour of linking PNs *via* a wordnet:

1. lexico-syntactic contexts which signal *instance of* links can be collected for many PNs and common nouns;
2. for various good reasons, PNs are already well represented in several wordnets.

As to argument 2: selected PNs are described in plWordNet because they are the derivational bases from which certain classes of frequent nouns and adjectives are derived, cf (Maziarz et al., 2011). Such PNs are part of the wordnet and are linked by plWordNet instance/type relations.

Argument 1 is even more important for us. We plan to describe semantically a very large number of PNs, and do it semi-automatically based on the information extracted from a large corpus (Kurc et al., 2013). Such information can support linking to a wordnet, but not directly to an ontology. Definite noun phrases are also used as anaphoric expressions to refer to and substitute PNs. Heads of such NPs are types for the substituted PNs or hypernyms of the proper types. That is yet another argument for linking PNs to an ontology via the wordnet as an intermediary.

A PN semantic lexicon will then be a separate resource linked to plWordNet 3.0 and through it to an ontology – more below. We will build up to 2.5 million Polish PNs an existing resource of 1.4 million.[12] The number of semantic categories will go from the present 52 up to more than 100. The categories will be mapped to plWordNet 3.0 synsets, providing a default link for each PN belonging to the given category. A more fine-grained mapping may be considered for selected categories such as persons. The PN lexicon is meant to be dynamic: it will be automatically expanded given any new corpus for a specific domain.

3.3 Wordnets and mapping

Unlike many other national wordnets constructed by the transfer and merge method, plWordNet has been built independently of PWN. That was a conscious choice motivated by the desire to offer a faithful description of a lexico-semantic system of Polish language, uninfluenced by the structure and

[12] See `http://nlp.pwr.wroc.pl/pl/narzedzia-i-zasoby/nelexicon`

content of PWN. Only when the core of plWordNet was constructed did we start its mapping to PWN (Rudnicka et al., 2012; Kędzia et al., 2013), noting a number of contrasts resulting from differences between lexical systems of English and Polish (*e.g.,* lexical gaps, lexicalised grammatical categories, different structuring of information) as well as in the content and structural design of the two networks.

The development of plWordNet 2.0 was independent of PWN (other than its evident influence as a general model). The mapping to PWN was manual, bottom-up, for selected domains – person, artefact, location, time, food and communication (Rudnicka et al., 2012). It was extended in plWordNet 2.1 to round out the coverage of those domains and to include PWN's core synsets (those representing the most frequent word senses) (Boyd-Graber et al., 2006). All this will facilitate linking to Open Multilingual Wordnet (Bond and Foster, 2013) and perhaps other similar resources.

The procedure considers several candidate inter-lingual relations (I-relations) in strict order. Initially, we placed inter-register I-synonymy – differently stylistically-marked words with close meaning – low on the decision list. It is, however, a well-defined choice when a marked Polish LU occurs in plWordNet but its counterpart is not in PWN, or even cannot be lexicalised in English. Now inter-register I-synonymy is next after I-synonymy. The same applies to inter-lingual partial synonymy, when there is a partial overlap of meaning and structure between the source and target synsets. The overlap is immediately visible, so partial synonymy can be assigned right after dismissing full synonymy. When neither I-synonymy applies, I-hyponymy is considered (it has turned out to be the most frequent I-relation), then I-hypernymy, I-meronymy and I-holonymy.

Manual mapping onto PWN is also an opportunity to verify plWordNet's content and structure, and repair errors. Linguists who did not create some part of plWordNet take a second look at it. The mapping procedure (Rudnicka et al., 2012) relies on the comparison of the relation structures for the corresponding synsets, so potential flaws in the hypernymy structure on either side can be discovered, especially because WordnetLoom visualises such structures (many levels down and up). The overall workload doubles in practice. Manual mapping takes nearly as long as wordnet construction, but if it includes verification then result is a lexical resource which allows a deep comparison of the two lexical resources on a very large scale.

The whole plWordNet 3.0 will be mapped onto PWN 3.1 (Rudnicka et al., 2012; Kędzia et al., 2013), and differences in lexical coverage will likely be a problem. A virtual supplement to Princeton WordNet 3.1 may be necessary to make the mapping work for Polish material not present yet on the English side (and give a boost to future multilingual applications). Gaps and discrepancies will be recorded and presented to the Princeton WordNet team. The mapping has thus far focussed on nouns. Extending it to verbs and adjectives may require a revised procedure.

3.4 The ontology

In plWordNet project we have deliberately kept the wordnet separate from any ontology, although we are aware that such a relationship must be established sooner or later. plWordNet has been built as a faithful description of the Polish lexical system providing an interface between the lexicon and abstract concept structures of an ontology.

Ontologies make concepts unambiguous, but natural language does not allow such "luxuries". Usage constrains meaning, and stylistic register is a case in point. Some lexical-semantic relations can link only words of identical or at least compatible registers.[13] Such considerations should be reflected in the wordnet structure. Constraints on registers in plWordNet 2.1 are part of the definitions of selected lexico-semantic relations: hyponymy and hypernymy can only connect words of compatible registers, inter-register synonymy accounts for near-synonymy with a tolerable register difference, and so on.

A wordnet's expressive power rests primarily on the lexico-semantic relations it encodes. One might say that, in the relational paradigm, all supplementary data, *e.g.,* glosses, are secondary, but such a strict position would yield wordnets inadequate for applications. Given that ontologies contain a different kind of information, it makes sense to create a mapping from a wordnet to an ontology and thus associate concepts with their lexical embodiment. Clearly, there is much linguistic knowledge not expressible by lexico-semantic relations, but it could appear in resources of other

[13] By way of illustration, two Polish words mean 'girl', but only *dziewczyna* is stylistically neutral, while *laska* is strongly marked as colloquial.

types linked to wordnets, such as syntactic and semantic valency frames (Hajnicz, 2012).

In theory, any ontology would work with plWordNet, but SUMO (Pease, 2011) ought to be favoured. There is a mapping from PWN (Peace and Fellbaum, 2010), and other wordnets linked to it are linked to SUMO at least indirectly. The manually constructed plWordNet-to-PWN mapping will help automate SUMO linking. I-synonymy links can be unambiguously mapped over. In other cases, ambiguity causes trouble, *e.g.,* between I-hypernymy and instances of SUMO hyponymy. Synsets in plWordNet and abstract SUMO concepts may have to be linked manually. The ontology mapping will enable the construction of an advanced shallow-semantic parser for Polish which builds a partial semantic representation from concepts acquired in SUMO via plWordNet. The ontology mapping will also facilitate linking plWordNet 3.0 to the Global WordNet Grid,[14] and will support the building of multilingual resources and applications.

4 The expectations

The construction of plWordNet 3.0 has started in July 2013. Complete plWordNet hypernymy branches are mapped to PWN in parallel by people other than those who built those branches. We expect plWordNet 3.0 to become a comprehensive wordnet (>200,000 lemmas) and one of the largest ever Polish dictionaries of any kind. The whole toolkit of semantic resources, completed by the end of 2015, will include plWordNet 3.0, a dynamic lexicon of 2.5 million PNs linked to plWordNet, a mapping plWordNet-PWN and a mapping of plWordNet to the top-level SUMO ontology plus selected medium-level ontologies. The lexico-syntactic structure of plWordNet MWEs (at least 60,000 lemmas) will be described in an associated resource. The toolkit will also be integrated with a syntactico-semantic valence lexicon.

The whole complex system of resources and tools (*e.g.,* for MWE and PN extraction), developed for the needs of the CLARIN project, is intended to be a strong, universal basis for applications and for further resources and tools, *e.g.,* a wordnet-based lexical similarity measure.

The modularly constructed toolkit will have a layered architecture of large software systems. Different layers of lexical knowledge will be separate but linked, *e.g.,* a relational description of lexical meaning in a wordnet and its formal interpretation in an ontology, or lexical meaning and facts represented by PNs. Each layer is based on limited set of notions and principles, can be used separately and upgraded.

Acknowledgments

Co-financed by the Polish Ministry of Education and Science, Project CLARIN-PL, and the Polish National Centre for Research and Development project SyNaT.

References

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

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

Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. 2006. Adding dense, weighted connections to WordNet. In *Proc. Third International WordNet Conf.*

Alan M. Collins and M. Ross Quillian. 1969. Retrieval Time from Semantic Memory. *Journal of Verbal Learning and Verbal Behavior*, 8(2):240–247.

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

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

Bradley Efron and Ronald Thisted. 1975. Estimating the Number of Unseen Species (How Many Words Did Shakespeare Know)? Technical report, Division of Biostatistics, Stanford University, California.

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

William A. Gale and Geoffrey Sampson. 1995. Good-Turing Frequency Estimation without Tears. *Journal of Quantitative Linguistics*, 2(3):217–237.

I. J. Good and G. H. Toulmin. 1956. The Number of New Species, and the Increase in Population Coverage, when a Sample is Increased. *Biometrika*, 43:45–63.

[14] See `http://globalwordnet.org/?page_id=67`

Jacob Grimm. 1999. *Deutsches Wörterbuch [The German Dictionary]*. Deutsche Taschenbuch Verlag.

Pierre Guiraud. 1954. *Les caractères statistiques du vocabulaire*. Presses Universitaires de France, Paris.

Elżbieta Hajnicz. 2012. Similarity-based Method of Detecting Diathesis Alternations in Semantic Valence Dictionary of Polish Verbs. In *Security and Intelligent Information Systems, SIIS 2011, Warsaw, Poland, Revised Selected Papers*, volume 7053 of *Lecture Notes in Computer Science*. Springer-Verlag. Pages 345–358.

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

Chu-Ren Huang, Nicoletta Calzolari, Aldo Gangemi, Alessandro Oltramari, and Laurent Prévot, editors. 2010. *Ontology and the Lexicon. A Natural Languge Processing Perspective*. Studies in Natural Languge Processing. Cambridge University Press.

Jan Karłowicz, Adam Antoni Kryński, and Władysław Niedźwiedzki, editors. 1900–1927. *Słownik języka polskiego [A dictionary of the Polish language]*. Warszawa.

András Kornai. 2002. How many words are there? *Glottometrics*, 4:61–86.

Ramesh Krishnamurthy. 2002. Corpus size for lexicography. Corpora list archive, ⟨http://torvald.aksis.uib.no/corpora/2002-3/0254.html⟩.

Roman Kurc, Maciej Piasecki, and Bartosz Broda. 2012. Constraint Based Description of Polish Multiword Expressions. In *Proc. Eight International Conf. on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. Pages 2408–2413.

Roman Kurc, Maciej Piasecki, and Stan Szpakowicz. 2013. Automatic Construction of a Dynamic Thesaurus for Proper Names. In A. Przepiórkowski et al., editor, *Computational Linguistics – Applications*, volume 467 of *Studies in Computational Intelligence*. Springer.

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

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

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, and Stan Szpakowicz. 2013a. Beyond the transfer-and-merge wordnet construction: plWordNet and a comparison with WordNet. In G. Angelova, K. Bontcheva, and R. Mitkov, editors, *Proc. Int.l Conf. on Recent Advances in Natural Language Processing*, Hissar, Bulgaria.

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

M. Nijhoff. 2001. *Woordenboek der Nederlandsche Taal [Dictionary of the Dutch Language]*. Instituut voor Nederlandse Lexicologie. First published in 1863.

Adam Peace and Christiane Fellbaum. 2010. Formal ontology as interlingua: the SUMO and WordNet linking project and Global WordNet. In Huang et al. (Huang et al., 2010).

Adam Pease. 2011. *Ontology - A Practical Guide*. Articulate Software Press, Angwin, CA.

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

Maciej Piasecki, Roman Kurc, Radosław Ramocki, and Bartosz Broda. 2012a. Lexical Activation Area Attachment Algorithm for Wordnet Expansion. In Allan Ramsay and Gennady Agre, editors, *Proc. 15th International Conf. on Artificial Intelligence: Methodology, Systems, Applications*, volume 7557 of *Lecture Notes in Computer Science*, Varna, Bulgaria. Springer. Pages 23–31.

Maciej Piasecki, Radosław Ramocki, and Marek Maziarz. 2012b. Automated Generation of Derivative Relations in the Wordnet Expansion Perspective. In *Proc. 6th Global Wordnet Conf.*, Matsue, Japan.

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

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

Laurent Prévot, Chu-Ren Huang, Nicoletta Calzolari, Aldo Gangemi, Alessandro Lenci, and Alessandro Oltramari, 2010. *Ontology and the lexicon: a multidisciplinary perspective*, chapter 1. In Huang et al. (Huang et al., 2010), pages 3–24.

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

Zygmunt Saloni, Marcin Woliński, Robert Wołosz, Włodzimierz Gruszczyński, and Danuta Skowrońska. 2012. *Słownik gramatyczny języka polskiego [A grammatical dictionary of Polish*. Warsaw University.

John Simpson. 2013. *Oxford English Dictionary*. Oxford University Press. ⟨`http://www.oed.com/`⟩.

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