Toward plWordNet 2.0

Maciej Piasecki
Institute of Informatics
Wrocław Univ. of Technology
maciej.piasecki@pwr.wroc.pl

Stanisław Szpakowicz
SITE, University of Ottawa &
ICS, Polish Academy of Sciences
szpak@site.uottawa.ca

Bartosz Broda
Institute of Informatics
Wrocław Univ. of Technology
bartosz.broda@pwr.wroc.pl

Abstract

Three years in development, the first release of a Polish wordnet contains almost 27000 lexical units grouped into some 17700 synsets. We look back at this completed first stage of the project, beginning with its main assumptions. We reconsider the challenges; we show how well they have been met and how many remain for the future. We discuss the benefits of semi-automatic wordnet expansion based only on information extracted from corpora. Next, we outline the plan for another three years, now devoted to growth – in the size of the network and in the variety of the underlying semantic relations.

1 All about plWordNet 1.0

There were no Polish wordnets or other NLP-friendly thesauri as late as in 2006[1]. The plWordNet project has set out three years ago to build such a resource[2] (Derwojedowa et al., 2009; Piasecki et al., 2009). Given very modest funding and the pressing need for a Polish wordnet, we assumed semi-automatic construction with a multi-purpose wordnet-builder's software tool and a significant input from semantic analysis algorithms. Such algorithms, notably to extract linguistic knowledge from large corpora of Polish, had to be designed and implemented first. We then constructed an integrated linguist's workbench – the WordNet Weaver (WNW) (Piasecki et al., 2009, section 4.5.3). It makes it much more efficient for the linguists to work on continual wordnet expansion, and helps increase the appropriateness of new entries. The WNW algorithm suggests additions – instances of wordnet relations – but the wordnet editors make all final decisions.

We wanted to make the first Polish wordnet suitable for a range of applications, not the least for linguistic research. It had to be an accurate description of the system of Polish lexical semantics, including its peculiarities. We followed the general idea of the Merge Model (Vossen, 2002) of wordnet development (Piasecki et al., 2009, Section 1.3.1), but the limited budget forced us to postpone mapping to other wordnets. We took a data-driven approach to the construction process: we based it on automated extraction of liguistic knowledge from large corpora of Polish. Details – see section 2 and (Piasecki et al., 2009).

With a clear focus on potential applications of plWordNet in language studies, we thought it consistent to put its structure on a solid linguistic footing. We started with lexico-semantic relations, zooming in on those usually featured in wordnets. Such relations link *lexical units*[3] (LUs) rather than synsets or other word groupings. This had led to the adoption of the LU as a type of vertex in the graph which underlies plWordNet – the basic building block of plWordNet's structure.

A team of linguists, assisted by the wordnet-builder's tool, manually created the first version of plWordNet with nearly 16500 LUs (Piasecki et al., 2009, Section 2.5). This was semi-automatically expanded to 26990 LUs in 17695 synsets (Piasecki et al., 2009, Section 5.2). The procedure can be summarised in 5 steps.

1. A set of candidate lemmas was collected from dictionaries and supplemented by the most frequent lemmas in a few large corpora of Polish (Piasecki et al., 2009, section 3.4.5).

2. Several data sets describing lexico-semantic relations were automatically extracted from morpho-syntactically tagged corpora. The extraction methods included measuring semantic relatedness and using lexico-syntactic patterns (Piasecki et al., 2009, chapters 2–3).

3. Selected groups of semantically close lemmas identified by a clustering algorithm

[1] There are two now; see (Vetulani et al., 2009).

[2] The Polish name is *Słowosieć*, literally word-net.

[3] A lexical unit is a word in a broad sense, possibly an idiom or a collocation, but not a productive syntactic structure. A lexical unit is represented by a lemma (a basic morphological form) and its meaning.

were loaded into WNW. The algorithm of activation-area attachment (Piasecki et al., 2009, section 4.5) suggested places in the plWordNet graph where LUs corresponding to the new lemmas might be added.

4. Linguists browsed suggestions in any order they found advisable, and edited the wordnet structure. (Reruns of the algorithm– triggered at will by editors – take into account the richer, enlarged wordnet structure).

5. A coordinator of the linguistic team reviewed the effects of such expansion, using the same WNW system.

In the remainder of the paper we will discuss how well we could keep the initial assumptions, what was helpful, what we had to change, and what we could recommend. We will also outline our plans for the recently begun next phase of the plWordNet project. We expect that its conclusion in three years will see a much larger plWordNet, with richer structures and built using improved methods of semi-automatic expansion.

2 Lessons learned

We applied the WordNet Weaver on a practical scale in the semi-automatic expansion of plWordNet from a manually-created *core* plWordNet to the present version 1.0. An effort of a mere 3.4 person-months resulted in the addition of 8316 new lemmas, 10537 new LUs, 8729 synsets and 11063 instances of lexico-semantic relations. The work performed encompassed also many improvements to the core plWordNet structure. It is our rough estimate that the expansion process was sped up 5–6 times in comparison to purely manual work. This estimate, however, is based only on observations made during an experiment whose goal was wordnet expansion. A systematic evaluation of the method is still to be performed. With the pressing goal of constructing a Polish wordnet as large as possible in the limited time, we could not afford working with the same lemma list independently in two different ways.

The semi-automatic approach means a process which is essentially data-driven. Extracted senses of lemmas were filtered by two factors: their coverage in the corpora and the ability of the extraction methods to recognize certain lexico-syntactic structures as clues during knowledge extraction. Even so, the editors always could – and occasionally did – override any suggestions and add missing LUs. The suggestions, however, often drew their attention to senses present in the data but not obvious; see (Piasecki et al., 2008) for a large collection of such cases. Here is an example of a word frequent in general language: *wigilia* ‘eve’ was also extracted as a sense of *święto* ‘holiday’. And an example from domain-specific language: *ocieplenie* ‘insulation’ was automatically linked to *izolacja* ‘insulation’, turned down by an editor but reinstated by the coordinator.

The activation-area attachment algorithm heavily relies on wordnet structure. An odd-looking, though consistent, suggestion may sometimes arise in the absence of hypernyms not entered yet into the plWordNet version under editing. The algorithm once suggested that certain types of food be linked to an incomplete, shallow hierarchy of *napój* ‘drink’. That happened because *potrawa* ’dish (food)’ and its hierarchy were absent in that version of plWordNet. The linking of *potrawa* to *napój* (as a hypernym/hyponym or synonym) was also suggested because the algorithm found it to be the closest match in that unfinished network.

New hyponyms were often associated with several co-hyponyms already present in plWordNet, because a link to a common hypernym was missing. As a result, several separate activation areas were suggested instead of one. For example, *pojazd mechaniczny* ‘mechanical vehicle’ appeared unconnected to many LUs denoting types of mechanical vehicles. Another example: for *bezpiecznik* ‘fuse’ the algorithm suggested three senses: *włącznik* ‘switch (for turning on)’, *przełącznik* ‘(change-over) switch’, *wyłącznik* ‘switch (circuit breaker)’. This was caused by the lack of a hypernym such as *element elektroniczny* ‘electronic element’. Such discoveries were a sure sign of likely mis-attachments or omissions in the network.

WNW’s focus on the hypernymy structure is twofold. First, hypernymy drives most of the search for a suitable location for new LUs,[4] so we lose structural information encoded in the instances of other relations. Second, WNW suggests only hypernymy links, while the generated suggestions relate a new LUs by other types of wordnet relation. We lose potentially valuable proper-

[4] One of the extraction algorithms (Piasecki et al., 2009, Chapter 4.5.1) is trained on data which include LU pairs associated by several wordnet relations.

ties if we present the suggestions to linguists only in terms of hypernymy and synonymy.

The hypernymy hierarchy of plWordNet is shallow for gerunds. A manual inspection of the WNW results gave an impression of suggestion accuracy[5] much lower for gerunds than for nouns. Yet, the accuracy is similar: 61.43% for gerunds or lemmas which are ambiguous between gerunds and nouns, 64.12% for lemmas unambiguously recognised as nouns (Piasecki et al., 2009, p. 169). The discrepancy may be due to the much higher number of suggestions generated on average per gerundial lemma than per noun. Most suggestions generated for gerunds rely only on the measure of (distributional) semantic relatedness, but evidence of different types increases the accuracy significantly. An effective description of gerunds requires different lexico-syntactic features. We return to this issue in Section 3.4.

Synonymy is an elusive phenomenon. There are many takes on it in linguistics. A synset is often described as a set of "words" referring to the same lexicalised concept (Fellbaum, 1998a; Vossen, 2002; Koeva, 2008). It is an even more demanding task to identify a concept, its extension and the corresponding set of "words" which lexicalise it. It seems easier to recognise, and consistently assign, lexico-semantic relations to pairs of LUs than to work with concepts. The analysis of LU meaning can be supported by usage examples found in corpora, while the connection between concepts and their uses can seldom be directly observed in language data.

The decision to make the LUs the centrepiece of plWordNet influenced its character a lot. First of all, the synset is no longer a primary element of the structure. A synset is defined via the lexico-semantic relations in which its members participate; it groups those LUs which share a set of lexico-semantic relation targets (Piasecki et al., 2009, section 2.1). A synset thus can be perceived as a *sui generis* "shortcut" for the fact that two or more LUs share the same relations. All structural links originate from the relations well studied in the linguistic tradition and the lexicographic practice. The centrality of LUs also allowed us to introduce substitution tests, similar to those applied in EWN (Vossen, 2002), in support of the linguists' work (Piasecki et al., 2009, Appendix A). The test were also automatically generated and available for every edit decision. Clearly, experienced editors could ignore the tests when making simple decisions, but tests instantiated with the appropriate lemmas were always at hand if needed.

We strive to base our definitions of lexico-semantic relations on facts observable in texts – such as those which substitution tests help uncover – and to avoid relying only on the editors' language competence. To support our decisions, we consulted with a few dictionaries and looked at material retrieved from large corpora. This stance leaves aside a variety of psychological and philosophical considerations around the issues of lexical meaning. We believe that such minimal commitment works to plWordNet's advantage: it is transparent to possible applications. For example, while plWordNet is not meant to be an ontology, it can be mapped to different ontologies if necessary.

Wordnet relations are usually categorised into those which link synsets and those which go between "words" or "literals". Making the LU a basic element of a wordnet makes this dichotomy unnecessary. All relations are defined at the level of LUs and are only inherited at the level of synsets.

The structure of plWordNet resembles – perhaps more closely than usual in wordnets – that of a monolingual dictionary with a dense network of lexico-semantic relations added. What makes plWordNet significantly different from a dictionary meant for human readers is the primary role of lexico-semantic relation; what makes it similar to a typical wordnet is the presence of synsets, even if they are defined in an unorthodox manner.

A consequence of our decisions is that plWordNet's hypernymy structure tends to be deeper (this was found out by manual inspection), while synsets are often smaller and represent a rather strict form of near-synonymy. For example, 79.5% of nominal synsets have only one lexical unit, 12.67% have two. A statistical comparison with Princeton WordNet (PWN) would be misleading so early in our project. WordNet's size and coverage are far above what we have achieved thus far: plWordNet only covers adequately the more general LUs and the upper levels of hypernymy. A selective manual comparison of plWordNet and PWN has revealed significant differences in the top hypernymy levels, but we have found no important differences beyond the discrepancy due to the obvious disparity of coverage and the rather

[5]We calculate it as the percentage of new lemmas for which the editors took up at least one suggestion.

predictable differences between the Polish and English lexical system, such as specific senses of some LUs. Our longer-term plans include a mapping between plWordNet and PWN, so we postponed a detailed, methodical comparison to that phase. Because plWordNet is developed bottom-up from extracted data, the top-level hypernymy hierarchy is mostly accumulated as new lemmas arrive and lower levels are linked. Only for some general lemmas in the initial version of plWordNet the upper hypernymy structure was created in advance. It is therefore difficult to draw any conclusions about the nature of divergences between the top levels of hypernymy in plWordNet and PWN.

We also statistically compared plWordNet with wordnets created during the first phase of the EWN project, with a similar duration of this phase. The average number of LUs in a synset is lower in plWordNet than in any other wordnet we analysed, but similar to the version of GermaNet from that period: 1.36 in plWordNet and 1.37 in GermaNet as reported in (Vossen et al., 1999).

3 Phase 2: a deeper and broader wordnet

3.1 Triple the Size

The results of the plWordNet 1.0 project are encouraging, but the wordnet is still small. While it has generated solid interest, it must grow considerably to became a real asset in NLP research and development. The WordNet Weaver, though already quite useful, must be extended in several ways. These are some of the themes of the follow-up project, funded for another three years.

Is there a *best* size of a wordnet? More is better, and in any event a large wordnet should at least match the size of PWN 3.0. That, however, costs time and money. Our *realistic* target for plWordNet 2.0 is 70000-80000 lexical units in 45000-55000 synsets. One of our objectives is to construct a mapping between plWordNet and PWN, as well some other wordnets (possibly in another project), we want to make plWordNet 2.0 comparable in size to contemporary large European wordnets, among them GermaNet[6] (Kunze and Lemnitzer, 2002).

In the plWordNet 1.0 project, we targeted 11 relations (Piasecki et al., 2009, Section 2.2), meant mostly to develop the nominal part of the wordnet. The follow-up project will put more resources into the verbal part and the adjectival part. When selecting new relations for inclusion in plWordNet 2.0, we will consider four main factors: lexico-semantic relations identified in linguistic study (especially for Polish), the requirements of plWordNet's expected applications of in Natural Language Engineering (NLE), the existing language resources for Polish (including those still in development), and relations described so far in wordnets (for compatibility).

In the following sections we describe three main research areas planned in the plWordNet 2.0 project. We focus on the inclusion of new lexical relations; a richer description of verbs in particular seems important for practical applications. On the other hand, we will also turn our attention to derivational relations, because derivational mechanisms are very productive and important in Slavic languages. Last but not least, we want to improve the methods of semi-automatic wordnet expansion for supporting linguists' work: as previously, we have little money and little time.

3.2 Enriching the Verbal and Adjectival Parts of plWordNet

In order to analyse quantitatively the description of LUs in terms of wordnet structural elements which deliver information about LUs, we have introduced the notion of *network density*. It is based on the number of relational instances – links – going from a LU to any other LU in the wordnet. Instances of symmetrical relations (such as antonymy) are treated as consisting of two links going in the opposite directions. In order to register the synset structure in the measure, two LUs occurring in one synset are treated as mutually linked. The network density in the verbal part of plWordNet 1.0 is high when we count synonymy (3.83), but quite low (1.52) if calculated without including synonymy represented by synsets; the density of coverage of nouns is 3.55 (2.48 without synonymy). The density corresponds to how well a LU is described on average by means of a set of different links, which defines the meaning of the LU according to the principles of a wordnet. Obviously the density is only a helpful factor and cannot be interpreted as an absolute measure.

Only several relations in plWordNet 1.0 can link verbal units[7]. For applications, it appears important to have a description of verb selectional

[6] www.sfs.uni-tuebingen.de/GermaNet/

[7] Those relations are hypernymy/hyponomy, troponomy, antonymy and conversion. We also considered some forms of derivation.

preferences in relation to verb subcategorisation frames. Such description, however, does not associate roles in verb frames with LUs. Rather, it considers semantic categories which can be defined as regions in the hypernymy structure. There is ongoing independent work on the construction of a Polish semantic valence dictionary (Hajnicz, 2009); role description in this work is based on plWordNet. We will thus focus on adding to plWordNet 2.0 several lexico-semantic relations based on the general notion of verbal entailment and modelled after the particular relations of this type used in PWN (Fellbaum, 1998b) and EuroWordNet (EWN) (Vossen, 2002). For example, we can consider *subevent* or *cause*; *manner* is already covered in plWordNet by troponymy.

For the Portuguese wordnet, WordNet.PT, Amaro (2006) proposed a set of much finer-grained verbal relations, motivated by the Generative Lexicon model (Pustejovsky, 1991). While the rich information expressed in that way could be valuable for NLE, such a set of relations would be very laborious to develop on a large scale and would make hard-to-meet demands on the automatic extraction of wordnet relations from text.

Initially only antonymy and derivational relations described adjectives in plWordNet 1.0. Grouping into synsets was based directly on near-synonymy (the construction of adjectival synsets was supported by the appropriate substitution tests (Piasecki et al., 2009, Apendix A). Adjectival hypernymy was added later in the project. The final network density is a low 2.72 (1.02 without synonymy), but this could be expected given the limited set of relations. Most solutions proposed in the literature for increasing network density would add cross-categorial relations to link adjectival LUs with nominal and verbal LUs; examples include *is a value of / attributes* known from PWN (Fellbaum, 1998c) or relations introduced in WordNet.PT (Marrafa and Mendes, 2006). Many such associations are expressed by derivational relations, so we want first to developed a semantic description of derivational relations, and only later to analyse those subsets of adjectival LUs which will still not be acceptably covered.

3.3 Towards a Rich Derivational Description

Derivational mechanisms, productive in Slavic languages, feature to some extent in all Slavic wordnets. Their builders – for example, Pala and Hlaváčková (2007) and Koeva (2008) – emphasize the importance of derivation. We also take this position (Derwojedowa et al., 2008; Piasecki et al., 2009). In plWordNet 1.0 derivational links appear as one of two relations: relatedness and pertainymy; the former covers more regular derivates. Irregular derivational LU pairs also belong to fuzzynymy. It is not enough, however, just to register the presence of a derivational link; its semantic status should be clarified. The idea is not new. Miller and Fellbaum (2003) introduced "morphosemantic links" into PWN to connect synset members. Fellbaum et al. (2009) proposed a semantic classification of types of associations expressed by the derivational relations.

Among the cross-categorial relations of EWN there are several whose instances in Polish would be associated by derivational links. Let us show example pairs of Polish words for the EWN relations. *Cross-categorial near synonymy*: *poruszyć*$_{verb}$ 'move [highly polysemous]' – *poruszenie* 'motion; commotion'. *Role/involved*: *wkręcać* 'screw in' – *wkrętak* '(a type of) screwdriver', *kierować* 'drive (a car)' – *kierowca* '(car) driver', *wypiekać* 'bake (bread, pastry)' – *wypieki*$_{plurale\ tantum}$ 'baked bread or pastry products'. *Co-role*: *pianino* '(upright) piano' – *pianista*$_{masculine}$ 'pianist', *gra* 'game' – *gracz* 'player', *sąd* 'court (of trial)' – *podsądny*$_{masculine}$ 'plaintiff'. *Be in state*: *biedak* 'poor or pitiful person' – *biedny*$_{adjective}$ 'poor'.

We want to follow this approach: adopt the classes for derivational relations proposed in (Fellbaum et al., 2009) to Polish and possibly extend the set, especially with NLE applications in mind. We will carefully examine the practice and experience of the EWN project (Vossen, 2002) in encoding cross-categorial relations and in their classification. For example, we will consider adopting for Polish the subtypes of the *role* relation[8]. Another interesting problem is the development of a semantic classification of aspectual pairs of verbs (such as *malować*$_{habitual}$ 'paint' – *pomalować*$_{perfective}$ 'have painted') in a way compatible with a relatively simple structure of a wordnet and conducive to NLE applications.

Derivational relations create in Polish (and in other Slavic languages) chains of associated LUs. An example: *bomba*$_{noun}$

[8] `agent`, `instrument`, `patient`, `location`, `direction`, `result`, `manner`, `source_direction`, `target_direction`.

'bomb', *bombardować*$_{verb}$ 'bomb', *bombowiec* 'bomber (plane)', *bombardier* 'bombardier (rank)', *bombowy*$_{adjective}$[9] 'related to bombs'... Another, rather breathtaking, example: *barwa*$_{noun}$ 'colour, hue', *barwić*$_{habitual}$ 'colour', *odbarwić*$_{perfective}$ 'remove colour', *zabarwić*$_{perfective}$ 'colour', *barwiarz*$_{masculine}$ 'dyer (person)', *barwiarka*$_{feminine}$ 'dyer (person)', *barwiarnia* 'dyeing shop', *barwny*$_{adjective}$ 'colourful', *pstrobarwny*$_{adjective}$ '$\approx$ motley'... Jadacka (1995) describes 64 (!) elements in a chain for *barwa*. It is a challenging task to make a semantic classification cover a wide range of indirect derivational association, while keeping the system's complexity in check.

We will introduce in plWordNet 2.0 a two-level model of the description of derivational relations between LUs: each link will be described by a pattern of a formal derivational dependency and a semantic type of the association. Pala and Hlaváčková (2007) constructed a tool for automatic prediction of derivational links between LUs from their word forms, and populated CzechWordNet with thousands of automatically generated derivational links. Since Czech and Polish are fairly close, a similar solution ought to be easily constructed for plWordNet; Rabiega-Wiśniewska (2006), for example, describes the lexical derivation in Polish with special focus on nouns and adjectives. Pala and Hlaváčková (2007) identified 10 productive derivational patterns, for example "action, verb $\rightarrow$ noun", and the produced links were labelled according to the pattern applied.

Derivational patterns often lead to ambiguous semantic classification of the LU links (Fellbaum et al., 2009; Koeva, 2008). For example, the affix *-ka* encodes different semantic relations in *miara* 'measure' – *miarka*$_{diminutive}$ 'measure' and in *krasnal*$_{masc.}$ 'brownie (dwarf)' – *krasnalka*$_{fem.}$ 'brownie (dwarf)'. LU pairs apparently related by derivation may have no semantic association, for example *mine* (bomb) – *mining* (industry) (Koeva, 2008) or *nakręcić* 'confuse (in a message)' – *nakrętka* '(threaded) nut'. We will deal with this difficulty by combining the pattern-based identification of derivationally linked lemma pairs with automatic extraction of information on the semantic nature of the association between the pair elements. We will rely on the positive experience with the combinations of different types of evidence extracted from corpora and classification methods based on Machine Learning (Piasecki et al., 2009, Section 4.5). WNW will be supplemented by a tool to suggest derivationally motivated links and their semantic classification.

Interestingly, the semantic ambiguity of derivational links is yet another good reason to base the wordnet structure on LUs. Ambiguity tends to disappear when we analyse semantic types of derivational relations at the LU level. For example, the semantic difference between the derivations *forma* 'form' – *foremny* 'shapely' and *forma* 'form' – *formalny* 'formal' is clarified when we know that the lemma *forma* represents two different LUs.

The introduction of new semantic relations via a semantic classification of derivational links creates a small dilemma. We should extend this description to LUs associated semantically in a similar way, but not linked derivationally. For example, *wykładać* 'to lecture' is derivationally related to *wykład* 'lecture'. The link can be seen as the `object` type, but this is also true of the pair *wykładać* – *przedmiot* 'subject matter', where there is no derivational relation. In plWordNet 1.0, pairs of the latter kind appear in the fuzzynymy relation. It is not a clear-cut decision whether to extend the semantic classification beyond derivational pairs and where to put the limits; associations of this kind have different strengths, are influenced by typicality, conventions, frequency of use and so on. For example, the definition of `involved/role` in EWN refers to typicality: "is typically involved" (Vossen, 2002, p. 29). This selection problem is not unlike difficulties with collocation, if one were to construct a dictionary of collocations. We will investigate all these matters for the needs of plWordNet 2.0 structure.

3.4 Deeper Research on Semi-automatic Methods of WordNet Expansion

There are at least two good reason why the plWordNet 2.0 project should expend much effort on the further development of extraction methods and on WNW: the limitations of the present version of WNW mentioned in Section 2 and the need to expand its operation to semantic classification of derivational links. WNW is based on several extraction methods delivering evidence for the final multi-criterial decision. Besides improving the accuracy and recall of the methods, we want to develop a uniform measure of the reliability of infor-

[9]... and an idiomatic meaning: great, fantastic, etc.

mation extracted by particular methods. All methods we used so far extract semantic associations between lemmas in the same part of speech. For the needs of semantic classification of derivational relations we will extend our methods to the description of cross-categorial pairs.

The description of nouns, in particular, is based on the occurrences of lexico-syntactic relations which characterise contexts of noun occurrences. Yet, three of four types of relations (Piasecki et al., 2009, p. 67) focus on those small parts of the sentences structure which include the noun (such as modification by a specific adjective) but not the verb. Only one type, "occurrence of *a specific verb* for which a given noun can be its subject" relates the noun to a verb. This limited view of syntactic relations was caused by the lack of a robust, shallow parser of Polish. A truly wide-coverage parser is yet to emerge. We will therefore continue work on morpho-syntactic constraints (Piasecki et al., 2009, Chapter 3.4.3) in order to collect a set of lexico-syntactic relations which can be recognised by the constraints with sufficient precision and describe lemma co-occurrences of more than one part of speech, for example verbs and nouns. The required changes are located mainly in the area of lexico-syntactic descriptions of occurrence contexts of the analysed lemmas. We will also develop a dedicated way of describing gerunds by adding morpho-syntactic constraints focused on the identification of possible elements of the gerund argument structure, motivated by derivation of gerunds from verbs.

WNW's algorithm of suggestion generation will be extended by taking into account links other than hypernymy. Additional evidence for the algorithm will come from derivational links already encoded in plWordNet and those automatically discovered when processing new lemmas. Cross-categorial information in the processing of verbal and adjectival lemmas could be the most reliable means of improving the performance of the next version of WNW with respect to lemmas in these two parts of speech. In parallel, WNW's User interface will be redesigned to give linguists properly visualisation of the whole wordnet structure. We want to make a flexible and scalable graphical presentation of the plWordNet structure the central element of the user interface, to set it as the default when presenting suggestions and editing the structure. List-based and record-based panels will be preserved to facilitate other kinds of tasks.

We will improve as well the mechanisms which support group cooperation and the management of the linguist team. We will build into the system access to external sources of knowledge, such as lexicons and encyclopaedias.

4 A few early conclusions

We have carried out a long and generally fruitful experiment in semi-automatic wordnet construction. We have made steady progress in developing methods of automated extraction of linguistic knowledge on a scale which makes them practically useful, and we deployed them with a most promising effect for a language significantly different than English. Our work also made good economic sense: a core Polish wordnet of some 16400 lexical units has been almost doubled in size fast and reliably (and that helped us make a convincing case for renewed funding).

We want to make plWordNet transparent with respect to theories of meaning and to applications in Natural Language Engineering. We hope to achieve it by reducing the methodological basis of the project to several linguistic notions such a lexical unit or a lexico-semantic relation. The present size of plWordNet is too low for drawing definite conclusions, but it seems to be developing in a good direction. It also is not as drastically different from other wordnets as we initially suspected it might become.

Methods of automatic extraction of lexical-semantic knowledge appeared to be mature enough for application by linguists in the practice of expanding the wordnet. The extracted data describe well the overall complexity of lexico-semantic relations, so the automatic support should be rather easily extended from hypernymy to many other types of relations.

The development of a wordnet in isolation from wordnets for other languages has several advantages (from our point of view – see Section 2). There is a risk: mapping to other wordnets and knowledge representation structures (perhaps some kind of general ontology) may not be an easy task. The separation of plWordNet construction and future mapping to other wordnets (or a general ontology) will most probably cost more than if both processes ran in parallel, but this choice of ours aimed at providing a faithful description of the Polish system of lexical meanings via plWord-

Net structures.

Acknowledgments

Work financed by the Polish Ministry of Education and Science, Project N N516 068637.

References

Raquel Amaro. 2006. WordNet as a Base Lexicon Model for the Computation of Verbal Predicates. In Petr Sojka, Key-Sun Choi, Christiane Fellbaum, and Piek Vossen, editors, *Proc. Third Global WordNet Conf.*, pages 9–17.

Magdalena Derwojedowa, Maciej Piasecki, Stanisław Szpakowicz, Magdalena Zawisławska, and Bartosz Broda. 2008. Words, Concepts and Relations in the Construction of Polish WordNet. In A. Tanács, D. Csendes, V. Vincze, Ch. Fellbaum, and P. Vossen, editors, *Proc. Fourth Global WordNet Conf.*, pages 162–177.

Magdalena Derwojedowa, Maria Głąbska, Maciej Piasecki, Joanna Rabiega-Wiśniewska, Stanisław Szpakowicz, and Magdalena Zawisławska. 2009. plWordNet 1.0 — The Polish Wordnet. Visit at `www.plwordnet.pwr.wroc.pl`, Apr. 09.

Christiane Fellbaum, Anne Osherson, and Peter E. Clark, 2009. *Putting Semantics into WordNet's "Morphosemantic" Links*, pages 350–358. In Vetulani and Uszkoreit (Vetulani and Uszkoreit, 2009).

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

Christiane Fellbaum, 1998b. *A Semantic Network of English Verbs*, chapter 3, pages 69–104. In (Fellbaum, 1998c).

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

Elżbieta Hajnicz. 2009. Problems with Pruning in Automatic Creation of Semantic Valence Dictionary for Polish. In V. Matoušek and P. Mautner, editors, *Proc. 12th International Conf. Text, Speech and Dialogue, Plzeň, Czech Republic, September 2009*, volume 5729 of *LNCS*, pages 131–138. Springer.

Hanna Jadacka. 1995. *Rzeczownik polski jako baza derywacyjna [The Polish noun as a basis for derivation]*. PWN, Warszawa.

Svetla Koeva. 2008. Derivational and Morphosemantic Relations in Bulgarian Wordnet. In Kłopotek et al. (Kłopotek et al., 2008), pages 359–368.

Claudia Kunze and Lothar Lemnitzer. 2002. GermaNet – representation, visualization, application. In *Proc. LREC 2002, main conference*, volume V, pages 1485–1491.

Mieczysław A. Kłopotek, Adam Przepiórkowski, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors. 2008. *Intelligent Information Systems XVI. Proc. International IIS'08 Conf., Zakopane, Poland, June 2008*. Advances in Soft Computing. Academic Publishing House EXIT, Warsaw.

Palmira Marrafa and Sara Mendes. 2006. Modeling Adjectives in Computational Relational Lexica. In *Proc. COLING/ACL 2006 Main Conf. Poster Sessions*, pages 555–562, Sydney, Australia.

George A. Miller and Christiane Fellbaum. 2003. Morphosemantic links in WordNet. *Traitement automatique de langue*, 44(2):69–80.

Karel Pala and Dana Hlaváčková. 2007. Derivational Relations in Czech WordNet. In *Proc. Workshop on Balto-Slavonic NLP*, pages 75–81, Prague, Czech Republic.

Maciej Piasecki, Bartosz Broda, and Michał Marcińczuk. 2008. The WordNet Weaver – support for semi-automatic wordnet expansion. Examples of suggestions automatically generated during the expansion of plWordNet, June 2008: `plwordnet.pwr.wroc.pl/browser/graphs.jsp`.

Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. 2009. *A Wordnet from the Ground Up*. Wrocław University of Technology Press. `www.site.uottawa.ca/~szpak/pub/A_Wordnet_from_the_Ground_Up.zip`.

James Pustejovsky. 1991. Generative Lexicon. *Computational Linguistics*, 17(4):409–441.

Joanna Rabiega-Wiśniewska. 2006. *A Formal Description of Lexical Derivation in Polish. Nouns and Adjectives* [in Polish[. Ph.D. thesis, Faculty of Polish Studies, Warsaw University.

Zygmunt Vetulani and Hans Uszkoreit, editors. 2009. *Human Language Technology. Challenges of the Information Society, Third Language and Technology Conf., Poznań, October 2007, Revised Selected Papers*. LNCS 5603. Springer.

Zygmunt Vetulani, Justyna Walkowska, Tomasz Obrębski, Jacek Marciniak, Paweł Konieczka, and Przemysław Rzepecki, 2009. *An Algorithm for Building Lexical Semantic Network and Its Application to PolNet - Polish WordNet Project*, pages 369–381. In Vetulani and Uszkoreit (Vetulani and Uszkoreit, 2009).

Piek Vossen, Claudia Kunze, Andreas Wagner, Karel Pala, Pavel Sevecek, Kadri Vider, Leho Paldre, Laurent Catherin, and Dominique Dutoit. 1999. Final WordNets for Czech, Estonian, French, and German. Deliverable 2D014, WP3, Wp4 LE4-8328, The EuroWordNet Project.

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