Approaching *plWordNet* 2.0

Marek Maziarz
Institute of Informatics
Wrocław Univ. of Technology
marek.maziarz@pwr.wroc.pl

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

Stan Szpakowicz
EECS, University of Ottawa &
ICS, Polish Academy of Sciences
szpak@eecs.uottawa.ca

Abstract

The Polish Wordnet, *plWordNet*, has been in steady development for five years. We are building it from scratch, all the time making provisions for its general compatibility with the other major wordnets. We are very close to reaching a milestone of 100000 lexical units in 70000 synsets. In addition to a fairly comprehensive coverage of common nouns, there already is in *plWordNet* a significantly built-up verb component, and a similarly enlarged adjective component is under construction. We present the background, the assumptions, the relation set (essential for any wordnet, and central for our inflection- and derivation-rich language) and the current state of the project, and we map the near future.

1 Promises

In 2009, we began the work on *plWordNet* 2.0, the next major release of the first large, publicly available Polish wordnet. The construction of *plWordNet*, initiated in 2005, has led to the release of a wordnet with 26990 lexical units in 17695 synsets (Piasecki et al., 2009, Section 5.2); *plWordNet* 2.0 has been planned (Piasecki et al., 2010b) as a very significant step toward a wide-coverage wordnet for Polish. We envisaged the expansion of *plWordNet* 1.0 in size and in the expressive power of the relation-based description. A special focus was to be given to verbs and adjectives, underrepresented in the 2009 release.

While no one can venture a guess at the ideal size of a wordnet, an optimistic target is to exceed the size of the largest existing dictionaries: a wordnet should describe lexical units which occur in textual data relevant to its numerous applications. The size of the *Princeton WordNet* (PWN) is still a hard-to-reach target for other wordnets. Our initial plans for *plWordNet* 2.0 were to make it comparable in size to what were then large European wordnets, among them *GermaNet*(Kunze and Lemnitzer, 2002), and to include all most frequent Polish lemmas. That meant 70000-80000 lexical units (LUs)[1] in 45000-55000 synsets. We revised our objective after receiving additional funds for the expansion of *plWordNet*: $\approx$ 135000 LUs in 90000-100000 synsets. We are already nearing 100000 LUs and 70000 synsets. The verb component is ready, and so is most of the noun component. Section 4 shows the detailed statistics.

The inventory of relations for nouns in *plWordNet* has been slightly modified; major changes are in place for verbs, and will be implemented for adjectives. Section 2 discusses the main assumptions and principles upon which *plWordNet* is founded. Section 3 briefly presents the system of relations. Section 4 presents the construction process. Finally, we discuss the experience gained and the work schedule for the last phase of the expansion.

2 Assumptions and Principles

PWN and most of other wordnets are structured into synsets. The synset is usually briefly described as capturing a lexicalised concept. A synset should contain a group of near-synonyms and represent the concept behind them, so that synset members share *some* meaning. How much is to be shared is left to the discretion of wordnet editors. An operational definition of synonymy is hard to formulate in a way which would support consistency of decisions among synset authors.

Synsets are linked by *conceptual relations* with names borrowed from linguistic work on lexical semantics, such as *hypernymy* or *meronymy*. Many lexico-semantic relations, however, clearly link LUs rather than sets of LUs (examples include various oppositions – including antonymy – and forms of derivation), and most wordnets note such links.[2] In fact, lexical semantics tends to say that hypernymy, meronymy etc. link pairs of LUs, not pairs of sets of LUs. We found it difficult to define simultaneously synonymy, synset and conceptual

[1] Without going into details, a lexical unit can be understood as a lemma with a sense number.

[2] See *lexical relations* (Fellbaum, 1998, p. 17) or *relations between word forms* (Miller et al., 1990), contrasted with *conceptual relations* or *relations between word meanings*.

relations. We proposed to adopt the LU rather than the synset as the centrepiece of the wordnet structure (Derwojedowa et al., 2008; Piasecki et al., 2009). Thus, lexico-semantic relations between LUs are primary, and from them we derive relations between synsets. LUs in a synset share certain (carefully selected) lexico-semantic relations: recognised by semanticists, well grounded in the wordnet tradition, frequent in language, with a reasonable *sharing factor*,[3] and with a potential to facilitate wordnet applications. They come with linguistically accurate *substitution tests* (Vossen, 2002; Piasecki et al., 2009), so a group of editors can annotate them consistently.

Synsets in *plWordNet*, then, are a notational convenience: to say that synsets S_1 and S_2 are linked by relation R is to say that any pair $s_1 \in S_1$ and $s_2 \in S_2$ is an instance of R. We refer as *constitutive relations* to the lexico-semantic relations selected to be the basis of synset construction. Different parts of speech require different sets of constitutive relations; see Section 3. Our experience has also shown convincingly that additional criteria are necessary to distinguish precisely between members of any two synsets. Such secondary factors include stylistic register for nouns, and semantic class and aspect for verbs (Maziarz et al., 2011a; Maziarz et al., 2011b).

The *plWordNet* project has always focussed on lexico-semantic facts specific to Polish. We decided to forgo the route which many wordnet developers take: translate PWN and adjust the result of that translation. Not only are we building the whole network (Piasecki et al., 2009), but we also design from scratch a system of relations to underpin *plWordNet*. Register and aspect (richly manifested in Polish) are among the prime consideration, though we constantly keep in mind the future – inevitable – alignment with PWN.

The hypernymy structure in PWN was initially a forest with *unique beginners*, only later joined into a tree. Potential links among very general LUs (such as *entity* or *abstraction*) are seldom well motivated by linguistic criteria. Not all abstract notions are lexicalised, so the introduction of *artificial lexical units* may be required. That is why in *plWordNet* we only introduce those hypernymy links which are compatible with the linguistic definition of hypernymy, and for which LU pairs pass the relevant substitution tests. This strategy must result in a hypernymy forest. On the other hand, *plWordNet* applications can only benefit from a single-root tree organisation (wordnet-based word-similarity calculation is a case in point). To meet application needs, we plan to map the top synsets (not linked by hypernymy or any other constitutive relation) onto a general, top-level ontology. SUMO (Niles and Pease, 2001) is among those considered.

We focus on the description of the Polish lexical system, so proper names (PNs) get very limited coverage in *plWordNet*. PNs are a very large, open category, which changes dynamically. Even if we wanted to select a limited subset, reasonably based on high frequency in a very large corpus, selection would be strongly biased by the origin of texts. We also wanted first to achieve nearly comprehensive coverage of "feasibly numerous" categories, mainly common nouns and verbs. Besides, the techniques of Named Entity Recognition support PNs and provide their classification. We made one exception: PNs which represent geographical objects and areas, and are the derivative bases for common nouns (such as inhabitants, for example "warszawianin" *Warsaw citizen* from "Warszawa" *Warsaw*) – a process very productive in Polish. Geographical names are a necessary completion of the description of the derivative common nouns.

In the relation-based paradigm of the lexical-semantic description, the number of relation links associated with a LU characterises well the amount of information encoded for this LU – each link adds to its differentiation from other LUs. In order to get good and balanced coverage of the description of LUs, we should aim at a wordnet which has at least several links for any LU. Piasecki et al. (2010b) propose to characterise this property by *network density*: the average number of relation instances – links – going from a LU to any other LU in the wordnet. Network density can be increased simply by increasing the number of relations, but an excessively detailed relation list would lead to an excessively fragmented description. The properties we postulate for the constitutive relations (Section 2) seem to be a good basis for selecting lexico-semantic relations for a wordnet. One constraint must be relaxed: we should not expect all wordnet relations to have high values of the sharing factor. For example, antonymy

[3]The sharing factor of a relation is the average size of a group of LUs which share this relation. Thus antonymy's sharing factor is 1 (a LU has at most one antonym) and hypernymy's usually well above 1.

is quite frequent but does not form LU groups.

At the early stages of the *plWordNet* 1.0 project there was no electronic dictionary on which we could base our work, so we adopted a corpus-based approach. The following recapitulation sums up a longer discussion in (Piasecki et al., 2009). First, lemma frequencies are generated from a very large corpus,[4] previously analysed morpho-syntactically, lemmatised and disambiguated. Next, proper nouns are filtered out, using a few large gazetteers (Marcińczuk and Piasecki, 2011). A morphological guesser is applied during the morpho-syntactic processing, so the list can include potential lemmas absent from the existing dictionaries. Lemmas with the highest frequency are selected if they are not yet in *plWordNet*. We usually take ≈ 9000 new lemmas in each phase of *plWordNet* expansion.

The corpus-based procedure allows us to include contemporary lexical units in the wordnet. In practice, however, every corpus is somewhat unbalanced, and that introduces a bias into the lemma frequency lists. That is why the process now includes consultation with dictionaries to correct flaws in corpus-derived frequency lists, though lemmas extracted from the corpus dominate. The reliance on the corpus imposes bottom-up direction on the construction of the wordnet hypernymy structure. There is no predefined hypernymy structure to import. Instead, LUs created for lemmas (extracted from the corpus) trigger the addition of synsets to link to the already existing hypernym synsets or to those recently added.

It is efficient, if linguistically not quite proper, to import language data from monolingual and bilingual dictionaries and existing wordnets. We could not, clearly, rely only on lemma frequencies and simple concordance. In a semi-automatic approach to wordnet development, we implemented several methods of extracting from our very large corpus potential instances of lexico-semantic relations. These raw data, combined in the *WordnetWeaver* system (Piasecki et al., 2009; Piasecki et al., 2011), suggest, for each new lemma, one or more LUs (Section 4). *WordnetWeaver* has been the main software support for the work of the editors, who still make all editing decisions but in a more efficient manner. Let us only note here that such partial automation complements the corpus-based development philosophy: automated tools ensure advanced semantic browsing and exploration of the language data and produce a condensed description of the discovered lexico-semantic dependencies for the editors.

[4]The present version contains 1.2 billion tokens taken from several publicly available Polish corpora – see (Piasecki et al., 2009) – plus texts collected from the Internet.

3 Relations

The system of relations in *plWordNet* has been fundamentally informed by the solutions in PWN and *EuroWordNet* (EWN), but also substantially influenced by the Polish linguistic tradition and the assumption that the lexical unit is the basic building block. Nine central relations have been defined for *plWordNet* 1.0 (Piasecki et al., 2009), not counting synonymy implicitly encoded in synsets; with subtypes of meronymy and holonymy the actual number was 19. Network density was relatively high for nouns, but too low for verbs and adjectives. There were also clearly fewer verb relations than in PWN and EWN. The rich Polish derivation was given in *plWordNet* 1.0 only two very general relations: *relatedness* and *pertainymy*. The *plWordNet* relation system is now much more involved: 17 relations among synsets and 16 among LUs, plus synonymy. With subtypes, there now are 44 synset and 42 LU relations. Many of them have a derivational character or originate from the derivational relations.

3.1 Synset relations

Synset relations are lexico-semantic relations extrapolated from the level of LU via the sharing of a relation between candidate synset members (Section 2). Substitution tests have been defined for each relation and relation subtype.

Hypernymy/hyponymy is defined for all parts of speech, only for LUs (extrapolated onto synsets) of the same part of speech. For nouns, the relation's definition is very similarly to that in *EuroWordNet*, see (Maziarz et al., 2011a). A handful of hypernymy/hyponymy instances in the adjective component of *plWordNet* 1.0 have yet to be revised. For verbs, we have decided to follow the practice of *plWordNet* 1.0, inspired by *EuroWordNet*, and refer as verb hypernymy/hyponymy to a special kind of entailment. The test was enriched with constraints which force both linked LUs to have the same aspect and belong to the same semantic verb class, see (Maziarz et al., 2011b).

Inter-register synonymy (defined between nouns and between verbs, considered for adjec-

tives) is «synonymy between lexical units which have different stylistic registers» (Maziarz et al., 2011a). It is used to link stylistically marked lexical units with their unmarked counterparts.

Holonymy/meronymy (nouns and verbs) is divided into subtypes. We have kept *part*, *place*, *portion*, *element of a collection* and *substance*, defined in *plWordNet* 1.0. A new subtype, *taxonomic unit*, expresses «lexico-semantic relations inside scientific taxonomies, especially biological taxonomy, for example, *kotowate* 'felidae' – *kotokształtne* 'feliformia'» (Maziarz et al., 2011a).

By analogy, holonymy/meronymy has been adopted for verbs. Two subtypes, *accompanying situation* and *sub-situation*, link verb LUs. The first «accounts for a 'primary' situation, represented by the holonym, typically supplemented by another situation, represented by the meronym» (Maziarz et al., 2011b). The second «associates a composite situation and its component», referring to a kind of temporal inclusion between the component and the whole (*ibid.*); it corresponds to the *subevent* relation in *GermaNet* (Kunze, 1999) and EWN (Vossen, 2002). For example, the verbs *trząść* 'shake [while travelling in a vehicle]' and *jechać* 'travel [in a vehicle]' are linked by *accompanying situation*, while *kryć* 'seek [in the hide-and-seek game]' and *bawić się w chowanego* 'play hide-and-seek' are connected by *sub-situation*. The two differ in that a typical situation of travelling in a vehicle *need not* be accompanied by shaking, whereas seeking *is* a typical (obligatory!) part of the hide-and-seek game.

Type/instance links synsets made up of proper names (synonymous names put into the same synset) to nouns which are their most specific descriptions. For example, ⟨*Wrocław*⟩ and ⟨*miasto* 'city'⟩ are linked by *type* relation. This is how it is done in *WordNet* (Miller and Hristea, 2006) and *EuroWordNet* (Vossen, 2002), where it is the relation *belongs_to_class / has_instance*.

Inhabitant (for nouns) arises from a specific but surprisingly productive derivational relation. Examples: *domownik* 'household member' – *dom* 'house', *wrocławianin* 'one living in Wrocław' – *Wrocław*. Because of the proper name variants and synonymous proper names, the relation was expanded beyond the derivational associations to link synsets, required to include the derivative and its base, respectively (Maziarz et al., 2011a).

The remaining relations, meant for verbs, are described in detail by Maziarz et al. (2011b). For full definitions and motivation please refer to that paper, from which we also took the « » quotations.

Cause (from verbs to verbs, nouns, adjectives) is a form of entailment, signalled in dictionary descriptions by verbs synonymous to "cause". The relation resembles *cause* relations in PWN (Fellbaum, 1998) and EWN (Vossen, 2002). There are two subtypes of *cause* for verb-to-verb pairs (pf-to-pf and impf-to-impf); four *cause of process* subtypes link verbs of different aspects with nouns and adjectives (pf-to-Adj, impf-to-Adj, pf-to-N, impf-to-N); and the *cause of state* subtype links perfective with imperfective verbs denoting states (pf-to-state), for example *uśpić* 'put to sleep$_{perf.}$' – *spać* 'sleep$_{imperf.}$'. This variety of subtypes, a little paradoxically, helps maintain coherence between editors: it simplifies test expressions.

Process (from verbs to nouns or adjectives) associates «verbs which denote spontaneous change of state or any dynamic situation» with nouns and adjectives describing the result of the change. The relation can be paraphrased using the verb *become*. It links synsets, but is often indicated by derivational associations, e.g., it links the synsets ⟨*chamieć* ≈ 'become$_{imperf.}$ a boor'⟩ and ⟨*prostak* 'simpleton', *cham* 'boor', *wieśniak* 'yokel'⟩. Four subtypes are defined by two values of aspect and two parts of speech (nouns and adjectives).

Inchoativity links verbs which describe either entering into a state or beginning an activity with verbs which describe *being* in this state or activity (in general, dynamic durative situations). There are two subtypes: perfective → imperfective and imperfective → imperfective verbs. An example is a link from ⟨*usypiać* 'put to sleep$_{impf}$', *zasypiać* 'fall asleep$_{impf}$'⟩ to ⟨*spać* 'sleep$_{impf}$'⟩.

State (from verbs to nouns or adjectives) expresses *being in a state*. It links stative verbs (representing static situations) with nouns or adjectives which describe a state. An example: *panować* 'rule$_{inf}$' is to be *pan* 'lord, ruler'.

Multiplicativity is a relation of a derivational character, with subtypes. *Iterativity impf-impf* «can link pairs of imperfective verbs such that one of them, which expresses an iterative meaning, is derived by suffixation from the other», for example, ⟨*pisywać 1* 'write$_{impf}$ sometimes'⟩ and ⟨*pisać 1* 'write$_{impf}$'⟩. *Iterativity impf-pf* subtype «can also link imperfective derivatives of *perfectiva tantum*»; for example, *zakochiwać się*

'fall$_{impf}$ in love sometimes' is the iterative form of *zakochać się* 'fall$_{pf}$ in love'. The third subtype, *distributivity*, associates a perfective verb which represents multiplicative performance of an action on many patients or by many agents with a perfective verb which denotes the performance of the whole process; for example, ⟨*nałowić* 'catch$_{pf}$ (plenty of)'⟩ and ⟨*złowić* 'catch$_{pf}$'⟩.

Presupposition «expresses the backward-going dependency between a situation represented by the given verb and a situation whose occurrence is a kind of precondition». The precondition is «mandatory regardless of the negative polarity of the sentence with the given verb». Example: ⟨*dawać* 'give'⟩ and ⟨*mieć* 'have'⟩.

Preceding is similar to *presupposition*, but the precondition is treated as desirable or holding in many, but not necessarily all, situations. Example: ⟨*popuścić 1* 'loosen'⟩ and ⟨*ścisnąć 1* 'press (together)', *zacisnąć 1* 'tighten'⟩.

All synset relations except inter-register synonymy are treated in *plWordNet* as constitutive: they meet the conditions defined in Section 2.

3.2 Lexical unit relations

Lexico-semantic relations not extrapolated to the level of synsets comprise two large groups: relations (motivated by linguistic or wordnet traditions) which do not express a sharing factor, and derivational relations. Synonymy is not directly described but it is encoded by synsets.

Antonymy applies to all parts of speech. It has been defined similarly to the definitions in PWN and EWN (Maziarz et al., 2011a), but divided into two subtypes: *complementarity* and *gradable opposition*. Complementary antonymy includes bipolar pairs of LUs with opposite and exclusive meanings, for example *man – woman*. Gradable antonymy links LUs with opposite senses which do not exhaust the semantic field, for example, *abstynent 1* 'teetotaller' – *pijak* 'drunkard'.

Converseness is a relation of oppositeness, applicable to nouns and verbs (Cruse, 1986, 10.6-10.7). It was considered in (Fellbaum, 1998), but in the end not included in PWN. For verbs, it is signalled by the mutually opposite roles assigned to the arguments, as in the classic pair *sell – buy*. Nouns are converses if they play opposite roles in some situation. A good substitution test is "If A is X (Prep) B, then B is Y (Prep) A" where X and Y are the nouns under investigation, such as X=*wife* and Y=*husband*.

Cross-categorial synonymy, always expressed by productive derivational patterns, has been defined for the noun-verb, noun-adjective and verb-adjective pairings. It has six subtypes, because the relationship is directional.

The **feature bearer** relation links a noun which represents an object characterised by some feature to an adjective which represents the feature, for example *starzec* 'an old man' – *stary* 'old'. **State** is an inverse of *feature bearer*.

Femininity links a feminine noun LU to its masculine derivational base. This relation, exemplified in English by *actress – actor*, is quite productive in Polish.

Markedness captures several forms of emotional markedness in derivationally associated nouns. There are three subtypes, all linking marked to unmarked words: *diminutive* (*armatka* 'small cannon' – *armata* 'cannon'), *augmentative* (*brzucho* 'big belly' – *brzuch* 'abdomen'), and *young being* (*wilczek* 'wolf cub' – *wilk* 'wolf').

Semantic role (noun to verb, as well as noun to noun) «characterises associations between a noun and derivationally linked verb from the perspective of a situation denoted by the verb» (Maziarz et al., 2011a). This relation is very similar to the *role* relation in EWN, and similarly subdivided into *agent*, *patient*, *instrument*, *location*, *product*, *time*, *agent of hidden predicate*, *object (of hidden predicate)* and *product (of hidden predicate)*. The last three subtypes are defined for derivationally associated pairs of nouns. The relation is directional: from a derivative to its base.

Role inclusion (verb to noun) is semantically opposite to **semantic role**, but the two relations are not mutually inverse. That is because they are always defined only for pairs: derivative and its base. The *role inclusion* subtypes are analogous to the first six subtypes of **semantic role**.

The **derivational** relation and **fuzzynymy** apply to all parts of speech. As in EWN, they are the last resort: the editor is convinced that two LUs are somehow related, but no regular *plWordNet* relation "works". Fuzzynymy is extrapolated to synsets by the relation-sharing rule.

4 The Construction Process

The growth of *plWordNet* from 27000 LUs to nearly 100000 LUs required the average workload of about 3.5 full-time editor positions over 1.5

years. It is notable that we did not resort to any form of translation from another wordnet, nor to importing data from any lexico-semantic resource. The construction process relied on the processing of a very large corpus. The editors could consult several dictionaries (Dubisz, 2004; Bańko, 2000) when they edited suggestions generated by the automated tools. We are confident that the fast pace of work was greatly assisted by the organisation of work we adopted, and by significant software support for the work of the linguists.

The work was divided into phases of 3-5 months. Each phase concentrated on the part of the network for one category (noun, verb, adjective). The first step was to firm up the definition of the relation system for this part of speech. This inevitably had wider consequences: there are many cross-categorial relations, so any change may affect relations for other parts of speech. Next, substitution tests are required for each relation and its subtypes. Tests – with a very strict structure – are treated as an intrinsic part of relation definitions. They are automatically instantiated with specific lemmas for testing, and systematically presented to the editor in a wordnet-editing system called *WordnetLoom* (Piasecki et al., 2009; Piasecki et al., 2010a). The number of relation subtypes had increased considerably because of the need to make test specifications formal.

Next, we select lemmas for addition to the wordnet and prepare knowledge sources which describe those lemmas for the automatic tools. Lemmas are extracted from our corpus.[4] During the first phases of *plWordNet* expansion, lemmas not recognised by the morphological analyser were filtered out; later we left on the list very frequent lemmas recognised by the morphological guesser. Next, we prune all proper names found in a large gazetteer (Marcińczuk and Piasecki, 2011). We always select 7000-9000 most frequent lemmas.

For the selected lemmas – combined with the lemmas MP already included in *plWordNet* – the following information is automatically produced from the corpus (Piasecki et al., 2009):

- Measure of Semantic Relatedness (MSR),
- lemma pairs extracted by hand-written lexico-syntactic patterns designed to detect hypernymy,
- lemma pairs extracted by automatically discovered statistical lexico-syntactic patterns,
- a classifier (trained on the data extracted from the corpus) designed to distinguish instances of *plWordNet* relation and other lemma pairs.

MSR, combined with the clustering system CLUTO (Karypis, 2002), groups the list of new lemmas into clusters of semantically related lemmas (50-200 in each). MSR and clustering introduce errors, and one lemma can represent several LUs, so flaws in the final clusters are inevitable. Nevertheless, each cluster represents about 2-3 different domains. Editors are next assigned clusters of lemmas to work on. This division of work, supported by *WordnetLoom*, enables them to concentrate on a limited number of semantic domains.

The extracted knowledge source are next delivered to the *WordnetWeaver*, a subsystem of *WordnetLoom*. For each new lemma, *WordnetWeaver* generates suggested LUs and presents them visually as subgraphs of the existing hypernymy structure. Editors are not limited by the suggestions: they can freely edit the wordnet.

There is even more support for editors, a recent addition to *WordnetWeaver*: automatically extracted examples of LU uses, produced by a system for unsupervised Word Sense Disambiguation (WSD) called *LexCSD* (Broda, 2011). *LexCSD* first identifies potential senses of a lemma by clustering its occurrences in the corpus. Next, for each cluster the most representative use is selected. For each new lemma, the generated examples are presented in the bottom part of the *WordnetWeaver* screen. While not all senses are automatically extracted, the size of the corpus (1.2 billion tokens) and the variety of texts and genres mean that the examples often include senses not covered by the existing dictionaries.

There were several stages of the expansion of *plWordNet* 1.0 toward 2.0. There were three stages devoted to nouns, considering that this is the category perhaps most important for potential applications of *plWordNet*. There were some 26000 new lemmas on the lists, but the final number of lemmas added was much higher. The editors included many synsets or even hypernymy subgraphs not on the extracted list; though frequency considerations dominate the expansion process, we decided that is better not to leave gaps in the new portions, because they might later be overlooked. The third stage saw some 9000 verb lemmas extracted from the corpus, some 13500 eventually added to *plWordNet* (more than 26500 new

LUs). The *plWordNet* statistics after the first three stages appear in Table 1. A new *plWordNet* is published every three months on the Web ⟨`www.plwordnet.pwr.wroc.pl`⟩ along with detailed statistics.

5 Lessons Learned

Semi-automatic wordnet-creation methods are far from producing results which would be acceptable without almost any human control. Nevertheless, they proved very useful: by "digging" into a very large corpus, they greatly helped increase the efficiency of the process and the coverage. It must be noted that a strict corpus-based procedure would almost certainly lead to many omissions clear to the native speaker. That is why we ask the editors to add units which they find obviously missing: a linguist, supported by a dictionary, can rather easily spot such lacunae.

The construction of the verb hypernymy structure benefitted from our verb classification. The verb class and the aspect are not elements of the relation-based description, but we refer to them in the definitions of relations. Both have influenced the relation system and became indirectly part of the description (Maziarz et al., 2011b).

The *plWordNet* structure is crucially shaped by the constitutive relations. They include no derivationally motivated relations, but relations which originate from derivational associations help differentiate LUs much more accurately. They link LUs, not word forms, and quite often only two particular LUs derived from the same lemmas are linked. As an example, consider the word "kometka". There are two LUs, *kometka 1* 'badminton' and *kometka 2* 'small comet', but the relation *markedness:diminutivity* can link only *kometka 1* to the LU *kometa 1* 'comet'.

6 More to Come

The present version of *plWordNet* is already large, but several expansion stages are still required to achieve the shape planned for version 2.0. First, we will create semi-automatically the hypernymy structure of nouns derived from verbs.[5] The structure will be based on the existing verb hypernymy structure. We want to add derivatives of the already described verb derivative bases. The analysis of a sample helped estimate that only 5% verbs will not have corresponding gerunds. Most of the verb hypernymy structure should be easily transferred to the noun component. The difficulty may be in merging the structures with the existing ones. Some gerunds were described in *plWordNet* 1.0. Also, verb hypernymy is more 'bushy', while gerundial structures will be mostly linked to the upper parts of the noun hypernymy structure. We expect to add some 20000 new noun LUs.

For the adjective component, a system of relations must be developed, perhaps inspired by a most interesting system in the Portuguese *WordNet.PT* (Marrafa and Mendes, 2006). We plan to add $\approx$ 15000 adjective LUs.

The Polish derivational mechanisms are relatively regular and very productive. We are working on automatic recognition of derivational relations with a tool trained on derivational pairs already described. The tool, applied to a long list of Polish lemmas, will identify derivatives and derivative bases not yet present in *plWordNet*. *WordnetWeaver* will be expanded to facilitate semi-automatic addition of LUs based the generated results. We expect to add 5000-8000 LUs.

The development of *plWordNet* has been monolingual all along, but mapping *plWordNet* synsets to *Princeton WordNet* synsets has always been an important element of our long-term plans. The process, slated for the year 2012, should provide mapping for at least 40000 of *plWordNet*'s noun LUs at the higher levels. We plan to perform the mapping in two iteratively repeated phases: first, verify and correct a selected part of the hypernymy structure (from the monolingual perspective), and then build a mapping for exactly the same hypernymy subgraph. We envisage some form of semi-automatic approach based on existing resources and methods. We expect that some new LUs can be added during verification and correction, so the final size of *plWordNet* at the end of the current project should reach 140000-150000 LUs in more than 100000 synsets.

[5] Polish deverbal nouns are similar to English gerunds, but they function more as independent nominal LUs.

Acknowledgments

Co-financed by the Polish Ministry of Education and Science, Project N N516 068637, and the European Innovative Economy Programme project POIG.01.01.02-14-013/09.

References

Mirosław Bańko, editor. 2000. *Inny słownik języka polskiego PWN*, volume 1-2. Wydawnictwo

	Nouns	Adjectives	Verbs	All
Lemmas	46746	3404	17009	67159
Lexical units	59467	4724	31133	95324
Synsets	44192	2791	21078	68061
Monosemous lemmas	37854	2582	9913	50349
Polysemous lemmas	8892	822	7096	16810

Table 1: *plWordNet* 1.5 in numbers, August 2011.

Naukowe PWN, Warszawa.

Bartosz Broda. 2011. Evaluating lexicographer controlled semi-automatic word sense disambiguation method in a large scale experiment. *Control and Cybernetics*, ??:??–?? (to appear).

D.A. Cruse. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge, UK.

Magdalena Derwojedowa, Stanisław Szpakowicz, Magdalena Zawisławska, and Maciej Piasecki. 2008. Lexical units as the centrepiece of a wordnet. In M. A. Kłopotek, A. Przepiórkowski, S. T. Wierzchoń, and K. Trojanowski, editors, *Proc. 16th Int. Conf. on Intelligent Information Systems*, pages 351–358.

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

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

George Karypis. 2002. CLUTO a clustering toolkit. Technical Report 02-017, Department of Computer Science, University of Minnesota.

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

C. Kunze. 1999. Semantics of verbs within GermaNet and EuroWordNet. In E. Kordoni, editor, *Workshop Proc. 11th European Summer School in Logic, Language and Information*, pages 189–200.

Michał Marcińczuk and Maciej Piasecki. 2011. Statistical proper name recognition in polish economic texts. *Control and Cybernetics*, ???:??? (to appear).

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.

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

Marek Maziarz, Maciej Piasecki, Stanisław Szpakowicz, Joanna Rabiega-Wiśniewska, and Bożena Hojka. 2011b. Semantic Relations Between Verbs in Polish Wordnet 2.0. *Cognitive Studies*, 11:183–200. `www.eecs.uottawa.ca/~szpak/pub/Maziarz_et_al_CS2011b.pdf`.

George A. Miller and Florentina Hristea. 2006. WordNet nouns: Classes and instances. *Computational Linguistics*, 32(1):1–3.

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

Ian Niles and Adam Pease. 2001. Towards a Standard Upper Ontology. In *Proc. Int. Conf. on Formal Ontology in Information Systems (FOIS-2001)*, pages 2–9, New York, NY. ACM.

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

Maciej Piasecki, Michał Marcińczuk, Adam Musiał, Radosław Ramocki, and Marek Maziarz. 2010a. WordnetLoom: a graph-based visual wordnet development framework. In *Proc. Int. Multiconf. on Computer Science and Information Technology - IMCSIT 2010, Wisła, Poland, October 2010*, pages 469–476.

Maciej Piasecki, Stan Szpakowicz, and Bartosz Broda. 2010b. Toward plWordNet 2.0. In P. Bhattacharyya, C. Fellbaum, and P. Vossen, editors, *Proc. 5th Global Wordnet Conf.*, pages 263–270. Narosa Publishing House.

Maciej Piasecki, Roman Kurc, and Bartosz Broda. 2011. Heterogeneous Knowledge Sources in Graph-based Expansion of the Polish Wordnet. In *Proc. 2nd Asian Conf. on Intelligent Inform. and Database Systems*, LNAI 6591. Springer. (to appear).

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