Words, Concepts and Relations in the Construction of Polish WordNet*

Magdalena Derwojedowa[1], Maciej Piasecki[2], Stanisław Szpakowicz[3,4], Magdalena Zawisławska[1], and Bartosz Broda[2]

[1] Institute of the Polish Language, Warsaw University,
{derwojed,zawisla}@uw.edu.pl
[2] Institute of Applied Informatics, Wrocław University of Technology,
{maciej.piasecki,bartosz.broda}@pwr.wroc.pl
[3] School of Information Technology and Engineering, University of Ottawa,
szpak@site.uottawa.ca
[4] Institute of Computer Science, Polish Academy of Sciences

Abstract. A Polish WordNet has been under construction for two years. We discuss the organisation of the project, the fundamental assumptions, the tools and the resources. We show how our work differs from that done on EuroWordNet and BalkaNet. In a year we expect the network to reach 20000 lexical units. Some 12000 entries will have been completed by hand. Work on others will be automated as far as possible; to that end, we have developed statistics-based semantic similarity functions and methods based on a form of chunking. The preliminary results show that at least semi-automated acquisition of relations is feasible, so that the lexicographers' work may be reduced to revision and approval.

1 Organisation of the project

Ever since the initial burst of popularity of the original WordNet [1, 2], there has been little doubt how useful *wordnets* are in Natural Language Processing. For those who work with a language that lacks a wordnet, the question is not *whether*, but *how* and *how fast* to construct such a lexical resource. The construction is costly, with the bulk of the cost due to the high linguistic workload. This appears to have been the case, in particular, in two multinational wordnet-building projects, EuroWordNet [3] and BalkaNet [4]. The recent developments in automatic acquisition of lexical-semantic relations suggest that the cost might be reduced. Our project to construct a Polish WordNet (plWordNet) explores this path as a supplement to a well-organized and well-supported effort of a team of linguists/lexicographers.

The three-year project started in November 2005. The Polish Ministry of Education and Science funds it with a very modest ca. 65000 euro (net). The stated main objective is the development of algorithms of automatic acquisition

* Work financed by the Polish Ministry of Education and Science, Project No. 3 T11C 018 29.

of lexical-semantic relations for Polish, but we envisage the manual, software-assisted creation of some 15000 to 20000 *lexical units*[5] (LUs) as an important side-effect. The evolving network also plays an essential role in the automated acquisition of relations. We describe the current state of the project in Section 3.3.

We will automate part of the development effort. A *core* of about 7000 LUs has been constructed completely manually; in a form of bootstrapping, the remainder of the initial plWordNet will be built semi-automatically. Algorithms that generate synonym suggestions from a large corpus [5] will make suggestions for the linguists to act upon. The ultimate responsibility for every entry rests with its authors, in keeping with our general principle of high trustworthiness of the resource. We must, however, try to reduce the linguists' workload and thus the time it takes to construct a network of a size comparable to several much more established European wordnets. We have allotted the funds approximately in the proportion 1 : 2 to manual work and to the software design and development work.

The remainder of the paper presents a more detailed overview of decisions made and work done till now, reviews the lessons learned, and sketches the plan for the last year of this project.

2 Fundamental assumptions

The backbone of any wordnet is its system of semantic relations. Two principles guided our design of the set of relations for *Polish WordNet* (plWordNet): we should — for obvious portability reasons — stay as close as possible to the Princeton *WordNet* (WN) set and the *EuroWordNet* (EWN) set, but we should also respect the specific properties of the Polish language, especially its very rich morphology. Tables 1 and 2 summarise our decisions[6].

In our description we have kept the division of lexemes into grammatical classes (parts of speech, as in WN): nouns, verbs and adjectives. Relations other than *relatedness* and *pertainymy* connect lexemes in the same class. Some relations are symmetrical (e.g., if A is an antonym of B, then B is an antonym of A; the hyponymy-hypernymy pair is symmetrical, too), while others are not (e.g., holonymy: a spoke is part of a wheel, but not every wheel has spokes). We refer to this property of semantic relations as *reversibility*.

In plWordNet, relations hold between LUs — pairs of lexemes. For example, the adjective *mądry* 'wise' is antonymous with *głupi* 'stupid', but its synonym *inteligentny* 'intelligent' has a different antonym, *nieinteligentny* 'unintelligent'; *mąż* 'husband' is a converse of *żona* 'wife', while its synonym *małżonek* 'spouse' has the converse *małżonka* 'spouse'. A derived form has obviously one root.

[5] We consider it a more precise measure of wordnet size than the number of synsets. Variously interconnected LUs – lexemes, generally speaking – are the basic building blocks of plWordNet.

[6] EWN has introduced a number of other relations which are not relevant to the discussion in this paper

WordNet	**EuroWordNet**	**Polish Wordnet**
synonymy	synonymy	synonymy
antonymy	antonymy	antonymy
		conversion
hypo-/hypernymy	hypo-/hypernymy	hypo-/hypernymy
mero-/holonymy	mero-/holonymy	mero-/holonymy
entailment		entailment
troponymy		troponymy
cause	caused/is caused by	
derived form	derived	–
pertainym	pertainymy	relatedness pertainymy
similar to		
participle		
see also		
attribute		
	role	
	has subevent	
	in manner of	
	be in state	
	fuzzynymy	fuzzynymy

Table 1. Semantic relations in WordNet, EuroWordNet and Polish WordNet

relation	**grammatical class**			**reversibility**
	noun	**verb**	**adjective**	
synonymy	+	+	+	+
hypo-/hypernymy	+	+	+	+
antonymy	+	+	+	+
conversion	+	+	+	+
mero-/holonymy	+	–	–	–
entailment	–	+	–	–
troponymy	–	+	–	–
relatedness	+	+	+	–
derived form	+	+	+	–
fuzzynymy	+	+	+	–

Table 2. Properties of the semantic relations in Polish WordNet

From EWN, we adopted the *fuzzynymy* relation. It is meant for pairs of lexemes which are clearly connected semantically, but which the lexicographer cannot fit into the existing system of more sharply delineated relations. The practice bore out our decision. We found, even in the basic vocabulary of the core list of lexical units, numerous instances of fuzzynymy (*przylądek* - *morze*, 'cape' - 'sea', *pacjent* - *przychodnia* 'patient' - 'walk-in clinic'). Future research includes a review of the fuzzynymy class to see if some subtypes of relations recur; this might be very interesting material for further linguistic investigation.

There is one relation unique to plWordNet: conversion (*narzeczony* - *narzeczona* 'fiancé' - 'fiancée', *rodzic* - *dziecko* 'parent' - 'child'), *kupić* - *sprzedać* 'to buy' - 'to sell'). Following Apresjan [6, pp. 242-265], we consider such cases to be different than antonymy.

Contrary to our initial expectation, hypo/hypernymy applies not only to nouns and verbs (*samochód* - *pojazd* 'car' - 'vehicle', *biec* - *poruszać się* 'to run' - 'to move'), but also to adjectives (*turkusowy* - *niebieski* 'turquoise' - 'blue'). In fact, adjectival hypo/hypernymy has turned out to be relatively widespread, once we allowed the lexicographers to note it.

Neither WN nor EWN support relations that enable an effective rendition of the semantic variation carried by rich morphology and productive derivation. In Polish, we have verb aspect (*szyć* - *uszyć* 'to sew - to have sewn'), reflexivity (*golić* - *golić się* 'to shave someone - to shave oneself'), subtle derivation via prefixes *(gnić* - *przegnić, nadgnić, wygnić etc.* 'to rot - to rot through - to become partially rotten - to rot out'), diminutives (*kot* 'cat' - *kotek, koteczek, kocio, kotuś, kotunio*; *mały* 'small' - *malutki, maluteńki, malusieńki, maluśki*), augmentatives (*dziewczyna* 'girl' - *dziewucha, dziewczynisko, dziewuszysko*), expressive names (*kobieta* 'woman' - *kobiecina* 'a simple or poor woman'), gender pairs (*malarz* - *malarka* 'painter$_{masc}$ - painter$_{fem}$'), names of offspring (*kot* 'cat' - *kocię* 'kitten'), names of action (*strzelać* 'to shoot' - *strzelanie* 'shooting', *strzelanina* 'fusillade'), names of abstracts (*nienawidzieć* 'to hate' - *nienawiść* 'hatred', *mądry* 'wise' - *mądrość* 'wisdom'), names of places (*jeść* 'to eat' - *jadalnia* 'dining room'), names of carriers of attribute (*rudy* 'red-haired' - *rudzielec* 'someone red-haired'), names of agents of action (*palić* 'smoke' - *palacz* 'smoker'), relational adjectives (*uniwersytet* 'university' - *uniwersytecki* 'university (in noun-noun compounds)'). Analogous phenomena were considered in Czech WordNet [7].

To account for this variety somehow, we decided to extend two relations, *relatedness* and *pertainymy*. In the former, we placed the most regular types of word formation: names of actions, abstract names, pure aspectual pairs (without any other semantic "surplus", e.g., *pisać* 'to write' - *napisać* 'to have written', *kupić* 'to have bought' - *kupować* 'to buy habitually or to be buying'), causative verbs (*martwić się* 'to worry' - *martwić (kogoś)* 'to worry someone'), relational adjectives and adjectival participles (which we do not consider as verb forms but as separate lexemes). The *pertainymy* relation accounts for the less regular word forms: names of places, carriers of attributes, agents of actions, offspring, augmentative, expressive and diminutive forms, gender pairs and names of na-

tionalities. The prefixed verbs and "impure" aspectual pairs are captured by *troponymy*. Although we tried to fit as much as possible into the WN and EWN relation structure, we agree with the Czech WordNet team: it is necessary to go beyond that set of relation if we are to take into consideration the specificity of Slavic languages (Pala and Smrž 2004; (86).

It is perhaps unexpected that the most problematic lexical-semantic relation turned out to be the fundamental one: synonymy. It helped little that this semantic notion is so well explored. There are two approaches to synonymy. One approach defines synonyms as lexemes with the same lexical meaning but with different shades of meaning; the other requires synonyms to be substitutable in some contexts [6, pp. 205-207]. In our opinion, neither approach works well in a semantically motivated network. We sharpened the criterion by positing that synonyms have the same hypernym and the same meronym (if they have any). For example, the lexemes *twarz, morda, gęba, ryj, pysk, facjata, buzia, pyszczek* (all of them mean more or less 'the face') can be considered synonymous in a wide sense. There are valid substitutions in some contexts (e.g., *dał mu w twarz/mordę/gębę* 'he hit him in the face'; *pogłaskała go po twarzy/buzi/pyszczku* 'she stroked his face'). They do not, however, have the same hypernym and meronym: *morda* is an expressive name of a face, but not a body part . We regard such expressive names as hyponyms of the unmarked lexemes such as 'face'; there is the same stance in [8]. One of the effects of this decision is that our synsets are very narrow, sometimes even with one element, but the hypo/hypernymy tree is much deeper.

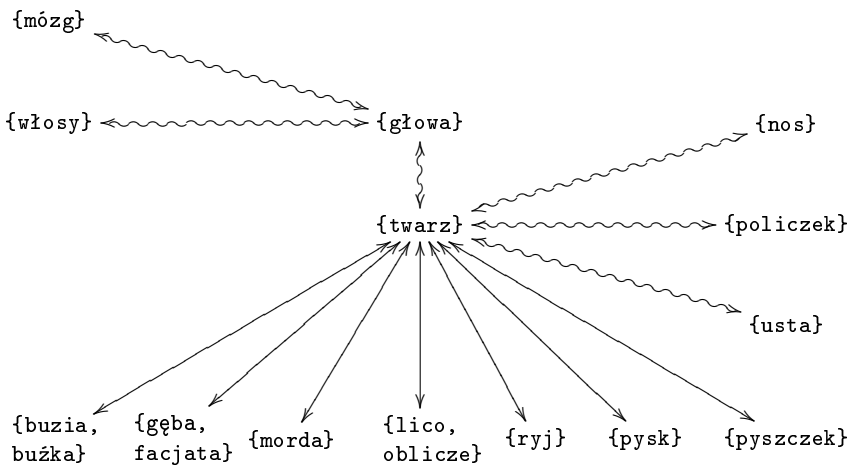

Fig. 1. The lexical unit *twarz* 'face' and its neighbours; straight arrows represent hypo/hypernymy, wavy arrows – meronymy/holonymy; *mózg* - 'brain', *włosy* - 'hair', *nos* - 'nose', *policzek* - 'cheek', *usta* - 'mouth'.

The problem with synonymy definition also arose in Bulgarian WordNet (Koeva, Mihev, Tinchov 2004; 62): "In Princeton WordNet the substitution criteria for SYNONYMY is mainly adopted [...] The consequences from such an approach are at least two — not only the exact SYNONYMY is included in the data base (a context is not every context). Second, it is easy to find contexts in which words are interchangeable, but still denoting different concepts (for example hypernyms and hyponyms), and there are many words which have similar meanings and by definition they are synonyms but are hardly interchangeable in any context due to different reasons — syntactic, stylistic, etc. (for example an obsolete and a common word)".

In our opinion, the vagueness of the synonymy definition and the lack of formal tools of establishing the synonymy of lexemes put in doubt the legitimacy of synonymy as the basic type of relation in lexical-semantics networks. It would appear that all relations link LUs. Suppose that B and D are (near-)synonyms and B is a hypernym of synonymous A and C; in certain contexts D may be substituted for B, and is also a hypernym of A and C.

The plWordNet project is building the semantic network from scratch; we decided not merely to translate the WN trees (in WordNet 3.0), because that would reflect the structure of English rather than Polish. We did try to translate the higher levels of WN, only to discover a few serious problems. 1) Many lexemes in WN can hardly be considered to denote frequent, basic or most general concepts in Polish; examples include *skin flick* 'film pornograficzny', *party favour* 'pamiątka z przyjęcia', *butt end* 'grubszy koniec', *apple jelly* 'galaretka jabłkowa'. 2) WN glosses are not precise enough to let us find the Polish equivalent, or there may be no lexical Polish equivalent at all (other than calques of English words); examples of untranslatable entries include *changer*, *modifier*, *communicator*, *acquirer*, *banshee*, *custard pie*, *marshmallow*. 3) Translating WN would create nodes in the hypo/hypernymy structure that represent unnecessary or artificial concepts; examples include *emotional person* 'osoba uczuciowa', *immune person* 'osoba uodporniona', *large person* 'duży człowiek', *rester* 'odpoczywający', *smiler* 'uśmiechający się', *states' rights* 'prawa stanowe'.

Our fundamental design decision was corroborated by the experience of the Czech WordNet team [7, pp. 84-85]. The BalkaNet project systematically recorded concepts from other languages (mainly from English, based on WN), not lexicalized in the language at hand. [...] The Czech team noticed problems with the translation of equivalents and the corresponding gaps with regard to English. They observed two types of cases where it was not possible to find synonyms (or even near-synonyms). The Czech synsets had no lexical equivalents in English because of the difference in lexicalizations and conceptualization, or because of the typological differences between those two languages;; there are, for example, no phenomena in English as the Czech as verb aspect, reflexive verbs or rich word formation. It is well known that concepts are not universal, nor are they expressed in the same way across languages (this is true even of so basic a notion as colour), although sometimes an ethnocentrism still can be observed — see Wierzbicka's criticism on that approach [9, pp. 193]. We decided to describe

the lexicalization and conceptualization in Polish as accurately as possible. We think that it is much more interesting to compare two semantic networks that reflect the real nature of two natural languages than to create a hybrid, which in fact would be just an English semantic network translated into Polish.

Near the end of year 2 of the plWordNet project, the noun network (the intended vocabulary) is ready. Work must be completed on verbs and adjectives. See Section 3.3 for more details.

3 Tools and resources

3.1 The linguist's tool

We now discuss software support for the Polish WordNet enterprise: a dedicated editor and algorithms that support lexicographers' decisions. Two years ago, all available tools – such as [10–12] – required editing the source format, not exactly linguist-friendly. A much more apt editor DEBVisDic [13] was not yet available[7]. We therefore chose to design our own wordnet editor, *plWNApp*, with tight coupling of the envisaged development procedure and the linguistic tasks. [14] present the implementation in some details; here, we focus on its use as a *tool*.

Linguists edit synsets and relations using plWNApp, which also supports verification and control by coordinators of the project's linguistic side. Written in Java, so practically fully portable, plWNApp has a client-server architecture with a central database. Clients *transparently* connect to the database via the Internet, though a version that allows work on a local copy of the database is also maintained. Efficiency, even on low-end computers, was a priority. Network communication is efficient due to caching data exchanged with the database. While it might put screen data out of synch for up to two minutes, this has not happened in 1.5 year of use by a large, distributed group of linguists.

Linguists work via a Graphical User Interface and never edit source files. Every user downloads an appropriate current version of the wordnet from the sever. Data are exported and archived in XML, in a special format that we plan to replace with a standard format once we have identified a fitting one. The coordinators can edit source files; they did that during the initial assignment of lexical units (LUs) to domains. The coordinators' stronger tool also supports definition of new lexical-semantic relations, invasive changes in the database and elements of group management. Both versions check on the fly such basic things as the existence of synsets/LUs or the appropriateness of relation instances to be added. More sophisticated diagnostic procedures have been designed, and some already installed.

Core plWordNet will have a complete description of selected LUs, so plWNApp distinguishes *system LUs* and *user LUs*. Only coordinators can add the former; other linguists introduce user LUs to complete synsets under construction.

Our linguistic assumptions suggested support for three main tasks:

[7] Early on, our project was also constrained by a commercial connection.

1. construct an initial, broad synset for a given system LU;
2. correct and divide initial synsets into more cohesive, almost always smaller synsets;
3. link synsets by lexical-semantic relations.

To support these tasks, plWNAPP's user interface features two *perspectives*: the *LU perspective* (2) and the *synset perspective* (3). The former is organised around selecting a LU and defining synsets and LU relations for it. A linguist would traverse the list of system LUs in the domain assigned to her and, for each LU, define all synsets to which it belongs. System LUs thus serve as starting points in synset construction.

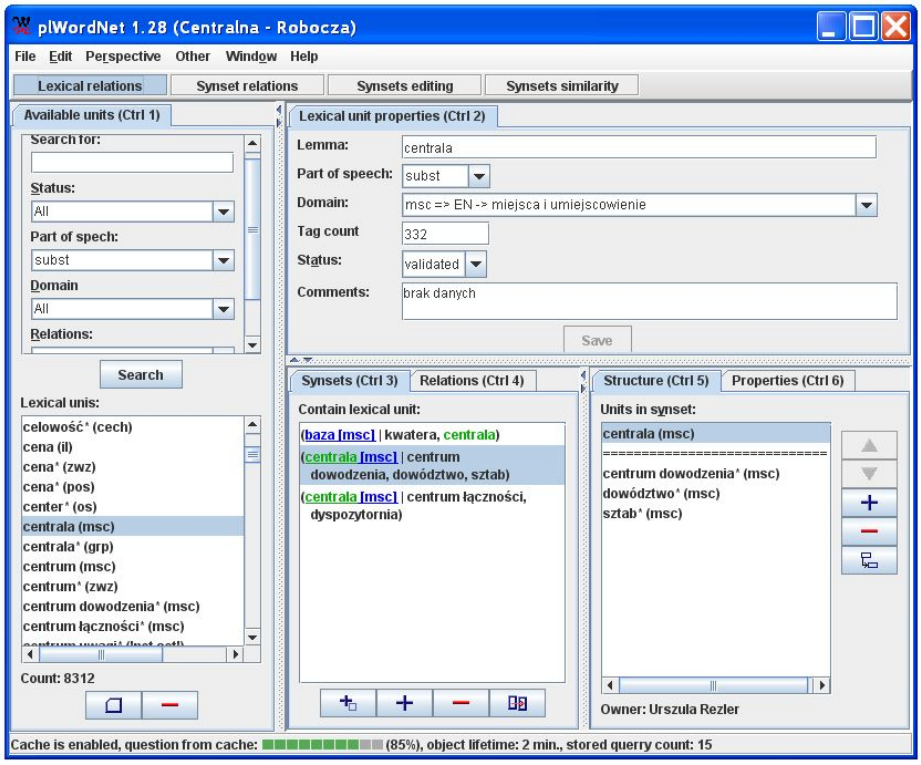

Fig. 2. The LU perspective

The intended result of task 1 was to group LUs in broad sets of near-synonyms, but pairs of synset often overlapped because of lack of precision in the grouping criteria. In order to support coordinators in task 2, we added the *comparison perspective*, showing two lists of synsets that share at least k LUs. Coordinators can edit or merge synsets, or move LUs around. We soon discovered, however, that correction – task 2 – is only possible when done together

with task 3, supported by the *synset perspective*. According to the definition of synsets and synset relations, a LU can participate in a synset only because of what we know about this synset's relations. In the *comparison perspective*, synsets are isolated from the structure of synset relations, and coordinators find it very hard to determine the correctness of the overlap between two synsets. In the next version of plWNApp, we will enhance this perspective to a comparison of structures of synset relations around two synsets.

In the *synset perspective*, each user interaction was to begin by the selection of a *source synset* which either must be corrected or is chosen as the starting node of a relation instance. Next, the user was to divide the source synset into two or to select a *target synset*, and then to pick a relation between the two (hypo/hypernymy when dividing the source synset). The added relation instances appear in a table at the bottom of the synset perspective. Predictably, practice diverged significantly from the initial ideas. The relation table was used most often, gradually becoming the central point of the synset perspective. Extracting a hypo/hypernym synset directly from the source synset was a very rare operation. Linguists preferred to create a new hypo/hypernym synset and move some LUs from the source synset, one by one. It may be easier to decide on one LUs than on a group. In any event, the synset perspective is the basic tool in transforming the initial synsets into the deepened hierarchy of narrow synsets, in keeping with our fundamental assumptions. Also, the table shows only relations of the selected source synsets, so linguists suggested extending the table to a graph view. We plan to introduce the possibility of editing synsets and synset relations in combination with the enhanced comparison perspective.

Early on, we found that consistency among linguists was a concern. In order to increase consistency, we introduced *substitution tests*. For each relation in plWordNet – for synsets and for LUs – there is a morphologically generic test with slots for LUs from the linked synsets or for the linked LUs. (Coordinators can edit definitions.) Slots are filled with the appropriate morphological forms. Whenever a relation instance is to be added, plWNApp generates a test instance and shows it to the linguist.

The tool associates domains not only with LUs but also with synsets. A LU is assigned to some domains when added it is to the database. The domain of a synset is that of its first LU, usually the system LU that started this synset. Domains offer a simple, but useful way dividing work among linguists. It is the coordinators' task to merge domain subsets. This is not trouble-free: occasionally, two linguists working on two close domains created a similar, overlapping structure of synsets and synset relation. An enhanced comparison perspective should help adjust such overlaps.

3.2 Toward automation

Work on extending plWNApp to support semi-automatic wordnet construction is under way. We will build software tools that:

- offer better corpus-browsing capability,

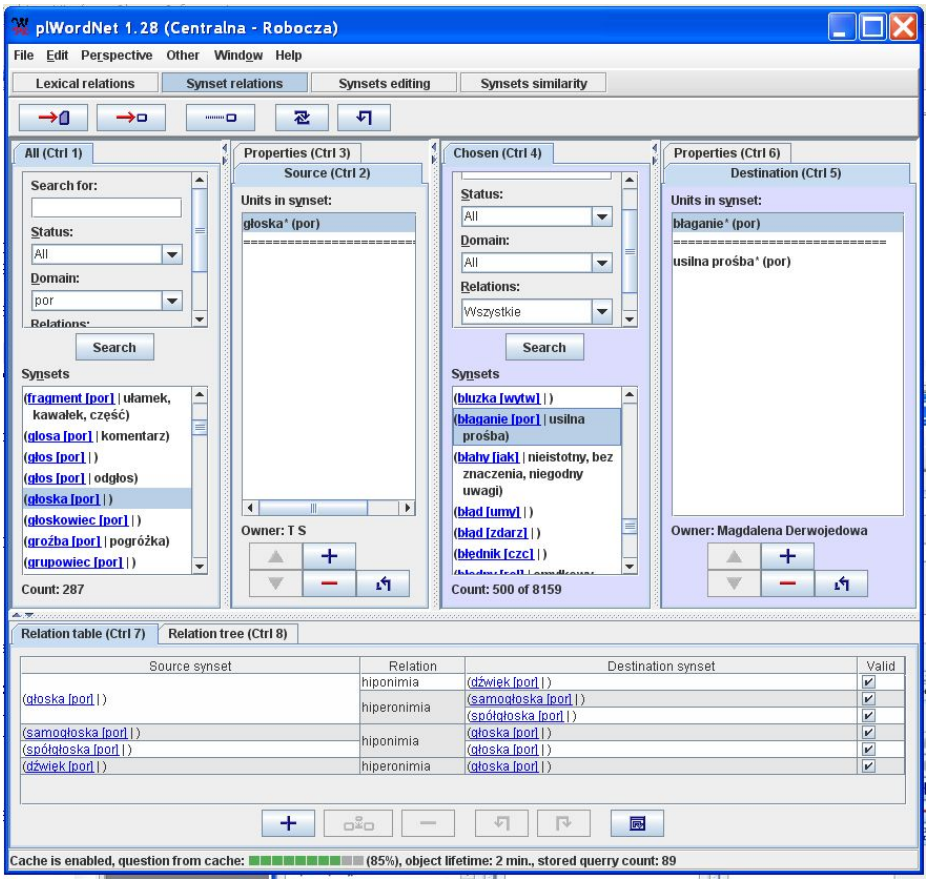

Fig. 3. The Synset perspective

– criticize existing wordnet content,
– suggest possible instances of relations.

The browsing tools are based on the statistical analysis of a large corpus in search for distributional associations of LUs. One can identify potential *collocations* and extract a *semantic similarity function* (SSF), which for a pair of LUs returns a real-valued measure of their similarity. As our examples showed, the real LUs are the minority among the extracted collocations, and it would be very hard to add new multiword LUs automatically on the basis of a collocation list. A linguist, however, can easily spot possible new multiword LUs if shown a candidate list.

SSFs are based on Harris's Distributional Hypothesis [15], aptly summarized in [16]: 'The distributional hypothesis is usually motivated by referring to the distributional methodology developed by Zellig Harris (1909-1992). (...) Harris' idea was that the members of the basic classes of these entities behave distribu-

tionally similarly, and therefore can be grouped according to their distributional behavior. As an example, if we discover that two linguistic entities, w1 and w2 , tend to have similar distributional properties, for example that they occur with the same other entity w3 , then we may posit the explanandum that w1 and w2 belong to the same linguistic class. Harris believed that it is possible to typologize the whole of language with respect to distributional behavior, and that such distributional accounts of linguistic phenomena are "complete without intrusion of other features such as history or meaning."'

Many methods of SSF construction have been proposed. The serious problem is their comparison. A SSF produces real values. Manual inspection of even several real numbers is very hard on people. While all known SSF algorithms produce interesting result, how do we choose a SSF that distinguishes really similar LUs (synonyms or close hypo/hypernym) from other groupings? Core plWordNet, constructed manually, can serve as the basis for evaluation. Following [17], we evaluate a SSF by applying it in solving a version of *WordNet-Based Synonymy Test* (WBST; see also [18]): given a word and four candidates, separate the actual synonym from distractors. The test is automatically generated from plWordNet; for evaluation, different SSFs were extracted from the IPI PAN corpus[8] [5] for the same set of LUs.

We tested several versions of SSF, achieving the best result of 90.92% in WBST generated from plWordNet for a SSF based on the Rank Weight Function (RWF) [19]: on the basis of SSF_{RWF} we can distinguish a synonym from three randomly selected words in some 90% cases. However, in a more difficult version of the WBST, called *Extended WBST* [18], in which decoys are chosen from LUs similar to the answer, the application of the same SSF_{RWF} gave the accuracy of 53.52% . Though the ability of SSF_{RWF} to distinguish among semantically related LUs is limited, it was added to plWNApp as a browsing tool. SSF_{RWF} is used to produce lists of LUs k most similar to the given one. Such a list can help linguists look among the top positions in the list for possibly omitted synonyms and hypo/hypernyms.

SSF_{RWF} is loosely correlated with similarity functions based on plWordNet but it is hard to find any threshold above which the similarity value guarantees the existence of the synonymy or hypo/hypernymy relation. In an experiment, we chose a value 0.2 as a threshold (on the basis of manual inspection). Next, one of the authors manually assessed a statistically significant sample of LU pairs with the similarity above the threshold, according to the synset relations: synonymy, hypo/hypernymy, meronymy and holonymy. Half of the pairs did not express any of these relations. The other half appeared to be worth browsing. In 7% of cases we found two synonyms already present in plWordNet, but only 1% of new synonym pairs. 20% of pairs were close hypo/hypernyms (not necessarily direct) already present in plWordNet, and 16% of new close hypo/hypernyms

[8] The IPI PAN Corpus contains about 254 millions of tokens and is rather unbalanced: most of the text in the corpus comes from newspapers, transcripts of parliamentary sessions and legal text, however also includes artistic prose and scientific texts.

and co-hyponyms were discovered. 1% of known meronyms and holonyms were found and 5% of new ones were discovered.

SSFs are intended to extract more rather than fewer semantic relations between LUs. We will reintroduce restrictions by way of clustering of the results of SSF – constructing proto-synsets. We also want to apply statistical lexico-syntactic patterns – for example, in the style of [20] – to a large corpus, in order to extract candidate instances of plWordNet relations. The extracted instances will be used to combine the cluster resulting from grouping LUs into a network of synset relations. The results of automatic extraction will be always anchored to plWordNet, because we want to extend it gradually, at each step adding a small set of new LUs automatically suggested for inclusion. After each iteration of automatic acquisition, linguists will be asked to verify and correct the proposed proto-synsets and instances of relations. The proposals will be clearly marked in plWNApp.

3.3 The current state of the system

At the time of this writing, plWordNet contains 12483 LUs grouped in 8095 synsets, 6059 synset relations and 5379 LU relations. Table 3 shows more detailed facts. While we feel that the number of LUs is more important than the number of synsets (Section 2), Table 3 separates relations between synsets and LUs — the former hold for every LU in a synset.

LUs		LU relations		synset relations	
nouns	8307	antonymy	1952	hypor/hypernymy	4293
verbs	3317	converse	47	holonymy	919
adjectives	3053	relatedness	1534	meronymy	847
		pertainymy	1175		
		fuzzynymy	671		
all	14677	all	5379	all	6059

Table 3. plWordNet in numbers, September 2007

The average rate of polysemy is 1.46 (calculated as the average number of synsets including the given homonymous LU, as in [1]), and the average size of a synset is 2.04 LUs. The detailed data appear in Tables 4 and 5.

4 Observations and future work

Our work to date has taught us a few valuable lessons. Of much use, though less interest, is what we found about facilitating the linguists' task. An important observation concerns the staring point of any properly conceived wordnet: it must be corpus-based. The core vocabulary should consist of words that are frequent

	Synsets to which a homonymous LU belongs											
	1	2	3	4	5	6	7	8	9	≥ 10	**Avg**	**WN**
All LUs [%]	73.45	16.10	5.98	2.32	1.02	0.58	0.30	0.09	0.05	0.12	1.47	–
Nouns LUs [%]	74.11	16.35	6.00	2.19	0.77	0.38	0.16	0.03	–	–	1.41	1.24
Verbs LUs [%]	79.40	14.73	4.04	1.34	0.28	0.18	0.04	–	–	–	1.29	2.17
Adj. LUs [%]	64.61	17.10	8.23	3.84	2.53	1.56	0.97	0.34	0.25	0.55	1.79	1.40

Table 4. The level of LU polysemy in plWordNet, September 2007 (WN means the Princeton WordNet 3.0)

	LUs in a synset									
	1	2	3	4	5	6	7	8	9	≥ 10
All synsets [%]	46.50	25.03	15.87	7.66	2.77	1.05	0.53	0.20	0.17	0.2
Noun synsets [%]	65.93	19.45	7.92	3.83	1.38	0.63	0.36	0.13	0.17	0.19
Verb synsets [%]	1.74	47.07	28.76	12.28	6.10	2.30	0.87	0.32	0.24	0.32
Adj. synsets [%]	15.69	26.17	33.04	17.29	4.87	1.47	0.87	0.33	0.13	0.13

Table 5. The number of LUs per synset in plWordNet

in real-life text. We have learnt that, for that particular purpose, certain balance in the corpus is extremely important. In our case, a little too many formal text resulted in a shortage of everyday vocabulary, such as names of the edible plants and food in general, animals and so on, in exchange for a higher than average number of economic and legal terms.

Experiments with translating the Princeton WordNet indiscriminately clearly show that only the top levels of the hierarchy may carry over to other languages intact; it probably *should* work, because this hierarchy may well be universal. We must work out the lower level afresh, if we want a wordnet that represents the lexical system, or at least much of the lexical system, of the langugage at hand — see Section 2.

Last but not least, we feel that for a wordnet to cover as much vocabulary of a given language as possible, it would need its own set of relations — many of them derivational in nature. This, however, would make it hard to use wordnets for multilingual NLP tasks, a most likely "killer app" of the near future. In the end, then, one ought to keep balance between too few but rather universal relations (such as antonymy or hypernymy) and too many, too detailed language-specific derivational relations. We believe that any criterion for choosing a useful set relations should consider the feasibility of future NLP tasks and linguistic credibility.

On the computing side of the plWordNet project, we see further fine-tuning of semantic similarity functions as a major task for the near future. Although the results thus far are very promising, too much noise can be observed in the data (about 50% — see Section 3.2). One cannot keep naive thresholds as a means of constraining the output of SSFs. We must first of all take a look at

multi-word expressions. We have already developed language-specific methods of extracting Polish multi-word expressions from a corpus, [21] but more work is neccessary. We need to build more natural groupings of words based on SSFs. One approach that we will try is to use fuzzy clustering algorithms. The preliminary results are again promising. On the other hand, pattern-based methods are very accurate and have been widely used to extract relations for WordNet; an early example is [22]). We will try to combine pattern-based methods with clustering. One way to accomplish this is to do machine learning of patterns on the basis of statistical and cluster information provided by a SSF; it should at least be useful in disambiguating lexico-semantic relations from the output of a SSF, but it also might help build the wordnet up in a weakly supervised manner.

References

1. Miller, G.A., Fellbaum, C., Tengi, R., Wolff, S., Wakefield, P., Langone, H., Haskell, B.: WordNet — a lexical database for the English language. Homepage of the project (2007)
2. Fellbaum, C., ed.: WordNet — An Electronic Lexical Database. The MIT Press (1998)
3. Vossen, P.: EuroWordNet general document version 3. Technical report, University of Amsterdam (2002)
4. Tufiş, D., Cristea, D., Stamou, S.: BalkaNet: Aims, methods, results and perspectives. a general overview. Romanian Journal of Information Science and Technology **7**(1–2) (2004) 9–43 Special Issue.
5. Przepiórkowski, A.: The IPI PAN Corpus: Preliminary version. Institute of Computer Science PAS (2004)
6. Apresjan, J.D.: Semantyka leksykalna. Synonimiczne środki języka (Lexical semantics. The means of synonymy in language). Ossolineum Wrocław (2000)
7. Pala, K., Smrž, P.: Building Czech Wordnet. Romanian Journal of Information Science and Technology **7**(1–2) (2004) 79–88
8. Dubisz, S., ed.: Uniwersalny słownik języka polskiego [Universal Dictionary of Polish Language], electronic version 0.1. PWN (2004)
9. Wierzbicka, A.: Język–umysł–kultura. PWN (2000)
10. Tengi, R.I.: 4. [2] 105–127
11. Louw, M.: Polaris user's guide. the EuroWordNet database editor. EuroWordNet (le-4003), deliverable d023d024. technical report, Lernout & Hauspie, Antwerp, Belgium (1998)
12. Horák, A., Smrž, P.: New features of wordnet editor VisDic. Romanian Journal of Information Science and Technology **7**(1–2) (2004) 201–213
13. Horák, A., Pala, K., Rambousek, A., Povolný, M.: DEBVisDic — first version of new client-server wordnet browsing and editing tool. In: Proceedings of the Third International WordNet Conference — GWC 2006, Masaryk University (2006) 325–328
14. Piasecki, M., Koczan, P.: Environment supporting construction of the Polish Wordnet. In Vetulani, Z., ed.: Proceedings of the 3rd Language and Technology Conference, 2007, Poznań. (2007) 519–523
15. Harris, Z.S.: Mathematical Structures of Language. Interscience Publishers, New York (1968)

16. Sahlgren, M.: The Word-Space Model. PhD thesis, Stockholm University (2006)
17. Freitag, D., Blume, M., Byrnes, J., Chow, E., Kapadia, S., Rohwer, R., Wang, Z.: New experiments in distributional representations of synonymy. In: Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005), Ann Arbor, Michigan, Association for Computational Linguistics (2005) 25–32
18. Piasecki, M., Szpakowicz, S., Broda, B.: Extended similarity test for the evaluation of semantic similarity functions. In Vetulani, Z., ed.: Proceedings of the 3rd Language and Technology Conference, 2007, Poznań. (2007) 104–108
19. Piasecki, M., Szpakowicz, S., Broda, B.: Automatic selection of heterogeneous syntactic features in semantic similarity of Polish nouns. In: Proceedings of the Text, Speech and Dialogue 2007 Conference. LNAI 4629, Springer (2006) 99–106
20. Pantel, P., Pennacchiotti, M.: Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, ACL (2006) 113–120
21. Broda, B., Derwojedowa, M., Piasecki, M.: Recognition of structured collocations in an inflective language. In: Proceedings of the International Multiconference on Computer Science and Information Technology — 2nd International Symposium Advances in Artificial Intelligence and Applications (AAIA'07). (2007) 247–256
22. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceeedings of COLING-92, Nantes, France, The Association for Computer Linguistics (1992) 539–545
23. Koeva, S., Mihov, S., Tinchev, T.: Bulgarian Wordnet ?– structure and validation. Romanian Journal of Information Science and Technology **7**(1–2) (2004) 61–78
24. Hamp, B., Feldweg, H.: GermaNet — a lexical-semantic net for German. In: Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid, ACL (1997) 9–15
25. Derwojedowa, M., Piasecki, M., Szpakowicz, S., Zawisławska, M.: Polish wordnet on a shoestring. In: Biannual Conference of the Society for Computational Linguistics and Language Technology, Tübingen. (2007) 169–178
26. Piasecki, M., team: Polish WordNet, the Web interface. (2007)