Maciej Piasecki, Stanisław Szpakowicz,
Bartosz Broda

A Wordnet from the Ground Up

Oficyna Wydawnicza Politechniki Wrocławskiej
Wrocław 2009

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

Reviewer
Tomasz Obrębski

Typesetting in LaTeX: Maciej Piasecki
(using LaTeX classes prepared by Adam Przepiórkowski and Marcin Woliński)

Cover design
Adam Radziszewski

Authors
Maciej Piasecki and Bartosz Broda
Institute of Informatics, Wrocław University of Technology, 50-370 Wrocław, Wyb. Wyspiańskiego 27
{maciej.piasecki,bartosz.broda}@pwr.wroc.pl
Stanisław Szpakowicz
School of Information Technology and Engineering, University of Ottawa;
800 King Edward Avenue, Ottawa, Ontario, K1N 6N5 Canada
Institute of Computer Science, Polish Academy of Sciences, ul. J.K. Ordona 21, 01-237 Warszawa, Polska
szpak@site.uottawa.ca

OFICYNA WYDAWNICZA POLITECHNIKI WROCŁAWSKIEJ
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław
http://www.oficyna.pwr.wroc.pl; e-mail: oficwyd@pwr.wroc.pl

ISBN 978-83-7493-476-3

Drukarnia Oficyny Wydawniczej Politechniki Wrocławskiej. Zam. nr 661/2009.

Contents

Preface

A language without a wordnet is at a severe disadvantage. If this sounds outlandish to you, reconsider. Language technology is a signature area of computing on/for/around the Internet, a growing source of texts for all manner of automated processing, including increasingly clever search engines and more and more adequate machine translation. A wordnet – a rich repository of knowledge about words – is a key element of many a successful text processing or language processing application. The English WordNet, whose origins date back almost a quarter century, is the exemplar. It has become central to much work in Natural Language Processing. Wordnets for other languages have been in development since the mid-1990s, and new projects start every year. We report on the initial stages of a long-term project to create a similar resource for Polish.

We have envisaged – though not quite achieved – a book for many audiences. The most immediate "clientele" are people who work with wordnets and on wordnets. We have attempted, without being too theoretical, to make our experience with one language approachable to people who need not know anything about that language.

Computing professionals who work with Polish texts may find the technical discussion interesting; we have presented a variety of tools which allow fairly deep analyses of meaning, given enough text to work with. Linguists who use computers in their study – and rely on well-organised language resources – may be encouraged to acquire yet another element of their research workbench. Researchers who specialise in statistical method of semantic analysis of texts may consider our comprehensive overview of such methods useful.

Chapters 3 and 4, which present our work on semantic analysis in substantial detail, are perhaps not as accessible as other parts of the book. Readers with little interest in these matters will not lose the main thread of the narrative if they only skim the two chapters.

There was no Polish wordnet when our work began several years ago. We chose to construct the resource from the ground up rather than translate the English WordNet first and then labouriously adapt it to the significantly different realities of the Polish language. A great team of linguists who have built the core of a Polish wordnet were assisted by a software tool designed and implemented by skillful programmers. Our

heartfelt thanks go to all those people[1]. Maria Głąbska was a particularly patient, thoughtful and diligent user of our software tools. A very special thank-you to our co-grantee Magdalena Zawisławska for her numerous contributions to the Polish wordnet, which she helped shape and jump-start. We are also very grateful to all those who offered generous advice and made their language tools and resources freely available to the plWordNet project[2].

A big special thank-you to Adam Radziszewski, who designed a handsome cover for our book.

The work which we present in the book has been financed by the Polish Ministry of Education and Science, project No. 3 T11C 018 29, and it has enjoyed the ungrudging support of the Wrocław University of Technology.

[1] Joanna Bilińska, Grzegorz Godlewski, Łukasz Jastrzębski, Paweł Koczan, Roman Kurc, Michał Marcińczuk, Dorota Lewandowska, Adam Radziszewski, Urszula Rezler, Anna Romańska, Ewa Rudnicka, Michał Stanek, Joanna Stankiewicz, Piotr Stawiński, Tomasz Stępień, Jakub Wilk, Beata Wójtowicz

[2] Aleksander Buczyński, Łukasz Dębowski, Filip Graliński, Elżbieta Hajnicz, Krzysztof Jassem, Klaudia Kunze, Lothar Lemnitzer, Karel Pala, Tadeusz Piotrowski, Jakub Piskorski, Adam Przepiórkowski, Zygmunt Saloni, Dawid Weiss, Marcin Woliński

Chapter 1

Motivation, Goals, Early Decisions

1.1 Motivation

1.1.1 What is a wordnet?

A wordnet is a computerised dictionary of synonyms, thesaurus, lexical database, taxonomy of concepts – the list can go on. Despite having been around nearly 20 years, wordnets still mean different things to different people – see the next section. People in the broad area of Computational Linguistics are quite familiar with wordnets, never mind the lack of consensus on a clear definition. We have come to count in research and in applications on the availability of such systems (Section 1.1.3). The first, and by far the best developed, among them is *the* WordNet (Miller et al., 2007) which we will refer to as the Princeton WordNet [henceforth PWN].

As a source of word senses, a wordnet resembles a thesaurus, and is often presented as a thesaurus (Fellbaum, 1998a, p. 210).

For NLP applications, a wordnet is an electronic resource that approximates the meaning of lexical units, though that is often limited to simple uses of hypernymy (a superclass-subclass relation). Synsets (groups of words closely related semantically) often merely supply alternatives or interchangeable additions to sets of keywords or search terms. There also have been, naturally, more imaginative applications that do justice to PWN's complex network of semantic connections. For example, *glosses* – informal definitions of senses represented by synsets – can be variously mined for some form of new knowledge.

1.1.2 Princeton WordNet

PWN is commonly used as a reference for other wordnets and for wordnet-related work. Using PWN as an exemplar, we will briefly analyse the various takes on the notion of a wordnet and the basic characteristics of a wordnet.

PWN began as a psychological experiment that aimed to explain how lexical meaning is stored in the mind, and to shed light on the acquisition of lexical meaning by children:

> WordNet is an on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. Miller et al. (1993, p. 1)

In the spirit of semantic networks, PWN is organised around abstract “lexicalised concepts” rather than around alphabetically sorted word forms or lexemes. There seem to have been no restricting assumptions on the notion of a lexicalised concept; see (Miller et al., 1993).

Vossen (2002, p. 5) proposed a similar description of and motivation for the notion of the synset and the sense of synonymy it expressed:

> A synset is a set of words with the same part-of-speech that can be interchanged in a certain context. [...] they can be used to refer to the same concept.

The nature and granularity of contexts is left to intuition[1].

One example should help clarify the intuition. There are seven senses of the noun *dog* in PWN (version 2.1 for Windows, 3.0 for Unix). Here are the corresponding synsets, ordered by estimated frequency:

Sense 1 {`dog, domestic dog, Canis familiaris`},

Sense 2 {`frump, dog`},

Sense 3 {`dog`},

Sense 4 {`cad, bounder, blackguard, dog, hound, heel`},

Sense 5 {`frank, frankfurter, hotdog, hot dog, dog, wiener, wienerwurst, weenie`},

Sense 6 {`pawl, detent, click, dog`},

Sense 7 {`andiron, firedog, dog, dog-iron`}.

PWN is often presented as organised around senses rather than lexemes. For example:

> Unlike a standard dictionary, WordNet does not take the word, or lexeme, as its elementary building block. Instead, WordNet resembles a thesaurus in that its units are concepts, lexicalized by one or more strings of letters, or word form. A group of words that can all refer to the same concept is dubbed a synonym set, or synset. [...] words and synsets are linked to other words and synsets by means of conceptual-semantic and lexical relations. (Fellbaum, 1998a, p. 210)

[1] In Section 2.1 we will return to the role of synonymy in the wordnet structure and to the nature of synsets.

An assortment of semantic relations other than synonymy hold between concepts represented by synsets. From the first release of PWN in the early 1990s, those relations were correlates of lexico-semantic relations (Section 2.2). In EuroWordnet [EWN] (Vossen, 2002) substitution tests proposed for detecting relations of this type operate on Lexical Units [LUs] (Section 2.1), not on synsets or lexicalised concepts.

The initial set of relations introduced in the early versions of PWN was extended in wordnets constructed later, such as EWN. We show several examples of relations and relation instances from PWN (version 2.1/3.0).

- For nouns:
  - *hypernymy* – {`tree, tree diagram`} is a kind of {`plane figure, two-dimensional figure`},
  - *hyponymy* – {`tree`} can for example be {`chestnut, chestnut tree`},
  - *holonymy* – {`mouth`} is part of {`face, human face`},
  - *meronymy* – {`mouth`} can have as its part {`dentition, teeth`};
- For verbs:
  - *hypernymy* – {`lollop`} is one way to {`walk`},
  - *troponymy* – {`lollop`} and {`stumble, falter, bumble`} are particular ways to {`walk`},
  - *entailment* – {`snore, saw wood, saw logs`} entails {`sleep, kip, slumber, log Z's, catch some Z's`},
  - *cause* – {`kill`} can cause someone to {`die, decease, perish, go, exit, pass away, expire, pass, kick the bucket, cash in one's chips, buy the farm, conk, give-up the ghost, drop dead, pop off, choke, croak, snuff it`}.

Some lexico-semantic relations were dropped between versions of PWN, e.g. the *similarity* relation for adjectives appears in PWN 1.5 but not in PWN 2.1/3.0.

Besides this evolution of relations in PWN, other wordnets also introduce changes and extensions. For example, in EWN the entailment relation has been divided into *has subevent* and *in manner*. Several relations linking parts of speech have been introduced. For example, *role* and *involve* (with many subtypes) link verbs with nouns that serve as instrument, agent and so on. *Cross-PoS near synonymy* marks the nominal synset {`motion, movement, move`} as a near-synonym of the verbal synset {`move`}. Instances of these relations are often derivational links between words that belong to the synsets; nonetheless, EWN formally defines such relations as semantic relations between synsets. Section 2.2.4 presents solutions adopted in plWordNet – a Polish wordnet introduced in Section 1.2 – against the background of the treatment of derivational relations in wordnets for other Slavic languages.

We note a discrepancy with all cited descriptions of a wordnet, which refer only to words as elements of synsets and the whole wordnet. It can be seen that in the PWN structure elements of synsets are LUs – one-word and multiword lexemes – represented by symbols built from a word and a sense number. Section 2.1 discusses the notion of the lexical unit in the context of plWordNet.

From the standpoint of linguistics, lexical relations hold between LUs, unmediated by sets of near-synonyms. PWN described three relations between words (LUs):

- *antonymy* (*tall* versus *short*, *wet* versus *dry*);
- two types of derivational relations:
  - *pertainymy* (the name adopted from EWN) defined only for relational adjectives (*sunny* pertains to the noun *sun*),
  - *related to* (the verb *paint* is related to the nouns *paint* and *painting*, the noun *run* is related to the adjective *runny*).

Antonymy is still a relation between words (LUs) in EWN (Vossen, 2002, p. 24), but its semantic variants – *near-antonymy* and *cross parts-of-speech near antonymy* – were also introduced as relations that link synsets based on the concepts those synsets represent. The set of derivation-based relations in EWN is identical to that in PWN. Wordnets for languages other than English, especially for Slavic languages – projects initiated between 15 and a few years ago – have adopted several lexical relations. We will discuss this issue in Section 2.2.4 that presents solutions adopted in plWordNet.

Let us return to the various informal definitions of a wordnet. Each of them has a practical rationale, dictated mainly by how PWN and then other wordnets have been applied in research and development (more on that in the next subsection). The freely available PWN 1.0 has been almost immediately picked up by people working in Natural Language Processing [NLP] as a resource that describes lexical semantics in the *relational paradigm*[2]. PWN is therefore often seen as "a large lexical database" (Fellbaum, 1998a, p. 209).

Even though it follows the principles of semantic networks – synsets represent lexicalised concepts and links represent conceptual relations – PWN is not a semantic network or an ontology in the sense accepted in Artificial Intelligence. A specialised notion of *lexical semantic network* has been proposed instead (Vossen, 2003). PWN has also been characterised as (Tufiş et al., 2004, p. 10)

[2] The meaning of lexeme L is described by the set of other lexemes that are in lexico-semantic relations with L.

> [a] special form of the traditional semantic networks [. . .] the concept of a lexical semantic network, the nodes of which represented sets of actual words of English sharing (in certain contexts) a common meaning[3].

PWN is also referred to as a *lexical ontology*. As Miller and Fellbaum (2007, p. 210) emphasise, however, PWN was never thought to be an ontology. PWN does include ontological relations – e.g., parts of the hypernymy hierarchy can be analysed as a taxonomy – but prior to the release of PWN 3.0 there was no distinction between *types* and *instances*. Even now most relations are linguistically motivated.

Our brief overview of the "takes" on the nature of PWN shows how important language intuitions are, especially for context-dependent synonymy. It all revolves around semantically motivated groupings of LUs. Surprisingly, the central building blocks of a wordnet's structure are typically not LUs but lexicalised concepts, about which few general assumptions are made. This gives a wordnet designer much freedom but precludes successful comparison, evaluation and especially interpretation of wordnets. The fast-growing number of "national" wordnets (more on that in Section 1.1.4) makes such inconsistencies problematic. It does not help that, with a few exceptions, those wordnets are new and rather small[4].

1.1.3 The importance of wordnets for language processing

For those who work with a natural language that lacks a wordnet, the question is not *whether* but *how* and *how fast* to construct such a lexical resource. The sheer number of applications and research experiments that rely on PWN (Fellbaum, 1998c) – just consider 868 projects listed in (Rosenzweig et al., 2007, state in Oct. 2008) – shows convincingly how useful wordnets are in NLP. Morato et al. (2004) presented a broad overview of the different PWN applications at the Second Global WordNet Conference; see the discussion of wordnet-related events at the end of this section. Rather predictably, the picture is not so clear when it comes to commercial applications, but PWN's free availability must have resulted in its inclusion in marketable products in the general area of language technology. Wordnets for other languages, even quite incomplete, are useful insofar as they are the only machine-tractable lexico-semantic resources for those languages.

The primary use of a wordnet may be as a *sense inventory*. For example, Agirre and Edmonds (2006, p. 7) characterise PWN as "the most-used general sense inventory in Word Sense Disambiguation research". Synsets, used as sense labels attached to words or expressions in text, help perform *Word Sense Disambiguation* [WSD] (Banerjee and Pedersen, 2002). Wordnet glosses are often used as a source of training data.

[3] In general, nodes in semantic networks may be labelled with abstract names.

[4] See `http://wordnet.princeton.edu/man/wnstats.7WN` for the current PWN statistics.

Semcor (Miller et al., 1993) is a part of the Brown Corpus (Francis and Kučera, 1982) annotated with PWN 1.6 word senses.

Wordnet-based WSD algorithms were applied – with mixed success – in Information Retrieval for semantic indexing: describe a document by word senses occurring in it and thus enable search by a comparison of the query meaning and the document meaning (Gonzalo et al., 1998, Moldovan and Mihalcea, 2000). Disambiguation is inherently hard in an inevitably short query of a few unrelated words (WSD usually depends on larger volumes of data). The limited coverage of wordnets is an issue likely to go away as those resources grow, also by means of robust semi-automated methods of wordnet construction. Wordnet-based WSD can work better in Information Extraction and Open Question Answering (Basili et al., 2002), where user queries tend to be complete sentences or syntactically rich phrases.

EWN (Vossen, 2002) and other aligned wordnets (see the next subsection) have much to offer to cross-language Information Retrieval and Information Extraction (Clough and Stevenson, 2004) and to Machine Translation (Dorr, 1997, Mohanty et al., 2008). A pair of aligned wordnets may deliver more helpful information than a traditional bilingual dictionary. PWN was applied also to the evaluation of the translation results (Agarwal and Lavie, 2008).

In Information Retrieval, the wordnet hypernymy structure can supply two mechanisms that facilitate query formulation: query narrowing (query term → its hyponym) and query broadening (term → hypernym) or more generally query expansion (Manning et al., 2008) – see for example(Moldovan and Mihalcea, 2000). The hypernymy structure has been useful in Information Extraction (Bagga et al., 1997), Open Question Answering (Clark et al., 2008) and Textual Entailment (Herrera et al., 2006). Hypernym pairs are the source of lexical chains applied in automatic Text Summarisation and Text Segmentation (Cramer and Finthammer, 2008).

Work on word similarity, word relatedness and analogy can benefit from the wordnet hierarchy. A range of methods have been proposed for computing semantic distance (called also terminological/conceptual distance or similarity) from the PWN structure (Budanitsky and Hirst, 2006). PWN-based semantic distance functions found applications, for example, in spelling correction (Hirst and Budanitsky, 2005), speech recognition (Pucher, 2007) and the processing of handwritten text (Zhuang and Zhu, 2005). Word classes based on hypernymy can help improve syntactic analysis and anaphora resolution. The PWN structure has been utilised in document structuring and categorisation (Fukumoto and Suzuki, 2001) and genre recognition (Klavans and Kan, 1998). There have also been approaches to the utilisation of the PWN hypernymy hierarchy as a kind of taxonomy in audio and video retrieval (Zaiane et al., 1999).

PWN inspired in some way the construction of several new semantic language resources. In addition to the semantically annotated corpus, Semcor, there is WordNet

Affect (Strapparava and Valitutti, 2004), an additional hierarchy of "affective domain labels" added to PWN, or WordNet Domains (Bentivogli et al., 2004), a grouping of PWN synsets and labelling them by domain. Other resources – FrameNet (Ruppenhofer et al., 2002) and PropBank (Palmer et al., 2005) are representative examples – have been motivated by the intention to address PWN's drawbacks or to add missing information (Miller and Fellbaum, 2007).

Finally, PWN has inspired research projects that aim to enrich the resource itself, e.g. manually created ratings describing the strength of association between two concepts (represented by two synsets) (Boyd-Graber et al., 2006) or the enrichment of PWN with folk knowledge and stereotypes (Veale and Hao, 2008).

The growing amount of research carried out on wordnets, based on wordnets and done around wordnets has inspired the organisation of the First Global Wordnet Conference [GWC] (Mysore, India) supported by the Global WordNet Association (GWA, 2008a) [GWA]. There ensued a series of successful biennial conferences. Workshops and sessions dedicated to wordnets and their applications take place at larger conferences. Seven events are listed on the GWA Web page (GWA, 2008a); others include the "Workshop on Usage of WordNet in Natural Language Processing Systems" during COLING/ACL'98 or "WordNet Special Track" during Language and Technology Conference in 2007.

Google Scholar[5] returned (on June 6, 2009) 4217 citations for (Fellbaum, 1998c) and 1336 for (Miller et al., 1990). There are thousands of citations to less well-referenced PWN-related papers.

1.1.4 Wordnets out there

The seminal project EuroWordnet (EWN) (Vossen, 2002) was initiated in 1996. The EWN project was aimed at developing wordnets for a number European languages, first Dutch, Italian and Spanish (PWN already covered English), and then Czech, Estonian, French and German. All wordnets were mutually aligned via the mediating mapping into *Inter-Lingual Index* introduced by the EWN project. Its records consist of an English synset, an English gloss that specifies the meaning and a reference to its source – to a synset in PWN 1.5. An upper-level ontology called *Top Ontology*, linked to Inter-Lingual Index, was also introduced in order to "to provide a common framework for the most important concepts in all the wordnets" (Vossen, 2002, p. 10). This orientation on the construction of aligned wordnets influenced the methods developed in EWN. We will return to this issue in Section 1.3.1.

The BalkaNet project (Tufiş et al., 2004) inherited the main assumptions and solutions from EWN. BalkaNet covered Bulgarian, Greek, Romanian, Serbian and Turkish,

[5] http://scholar.google.com

as well as Czech (that wordnet was expanded from the state in EWN). The first ever multi-lingual wordnet project was CoreNet (Choi and Bae, 2004) – a Korean-Chinese-Japanese initiative that linked the three languages via a hierarchy of shared semantic categories. It began in 1994. There is high potential for multi-lingual wordnets in NLP applications. New projects appear (Sinha et al., 2006). More on that at the end of this section.

EWN focussed on European languages. Unilingual wordnet-construction crop up all across the world. Let us just list a Hindi WordNet (Debasri et al., 2002), a Farsi wordnet FarsNet (Shamsfard, 2008), African WordNet (Le Roux et al., 2008) and Arabic WordNet (Rodríguez et al., 2008). Wordnets for European languages under development after the completion of the EWN project and outside the BalkaNet project include a Danish wordnet DanNet (Pedersen and Nimb, 2008), a Hungarian WordNet (Miháltz et al., 2008) and a Russian wordnet RusNet (Azarowa, 2008),

More than one wordnet is in the works for some languages, including Bulgarian, Korean and Russian (GWA, 2008b). A Polish project whose one of the goals is the construction of a wordnet called PolNet (Vetulani et al., 2007) started a year after the plWordNet project. PolNet is being constructed with particular applications in a homeland security system in mind.

In October 2008, 65 wordnets were listed on the web page (GWA, 2008b) maintained by the Global WordNet Association (GWA, 2008a).

There is a growing number of wordnets – lexico-semantic language resources that follow a similar blueprint – many of which are inter-related directly or via PWN. This observation prompted the idea of a global system of lexical networks. Proposed at the third Global WordNet Conference in Korea (January 2006) (GWA, 2008a), it is actively promoted by GWA (Fellbaum and Vossen, 2007) – note, too, Piek Vossen's invited talk at the LTC'07 conference in Poznań (Vetulani, 2007). The relations between individual wordnets can be complex due to the idiosyncratic linguistic properties of languages with different typologies. The proposed solution, called the Global WordNet Grid, is to be based on anchoring of the many of existing wordnets to a shared ontology with some 5000 shared common concepts. The ontology will be derived from the Suggested Upper Merged Ontology [SUMO] (Niles and Pease, 2001).

The Global WordNet Grid is meant initially as a language resource that supports various applications of language technology, for example Information Retrieval and content mining from language documents in various languages in order to map information and knowledge expressed. The idea of the Global WordNet Grid has been implemented, for example, in the KYOTO project (Knowledge-Yielding Ontologies for Transition-Based Organisation). Its main goal is to develop "a content enabling system that provides deep semantic search" (Vossen et al., 2008, p. 475). Processing of multimedia data expressed in several languages, including European and non-European languages, is intended. The search will be based on dialogue with the user, and the domain is limited to the natural environment and ecology. The very important characteristic feature of the Global WordNet Grid and the KYOTO results is the assumption of their availability via a form of free public licence.

1.2 The Goals of the plWordNet Project

Till 2005 there have been no wordnets or other NLP-friendly thesauri for Polish. To remedy this deficiency has become urgent in view of the overall importance in NLP of PWN and to some degree other wordnets – see Section 1.1.3. The building of a large wordnet for Polish is the main objective of our long-term research agenda. The timing prevented our participation in a large international project such as EWN or BalkaNet. On the other hand, we were free to construct a trustworthy resource – an essential characteristic of every wordnet – motivated in every detail by the considerations relevant specifically to the properties of the Polish language.

This book sums up our experience, which seems quite different from the experience of many other wordnet projects, and presents the design and development process with all its potential positives and negatives.

The construction of a wordnet is costly, with the bulk of the cost due to the high linguistic workload – see the discussion in Section 3.1. This appears to have been the case, in particular, in two multinational wordnet-building projects, EWN (Vossen, 2002) and BalkaNet (Tufiş et al., 2004). The recent developments in automatic acquisition of lexico-semantic relations suggest that the cost might be reduced. Our project to construct a Polish wordnet explores this path as a supplement to a well organized and supported effort of a team of linguists.

The three-year project started in November 2005. The Polish Ministry of Education and Science has funded it with a very modest $\approx$ 65000 euro (net). The stated main objective was the development of algorithms of automatic acquisition of lexico-semantic relations for Polish, but we envisaged the manual, software-assisted creation of some 15000 to 20000 LUs[6] 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 5.2. We named the constructed wordnet system *plWordNet*[7] (Derwojedowa et al., 2008).

[6] We consider the number of LUs described in detail as a more precise measure of wordnet size than the number of synsets. We argue in Section 2.1 that variously interconnected LUs are the basic building blocks of our wordnet. The number of LUs described also gives a clearer information of the wordnet coverage for NLP applications.

[7] The Polish name *Słowosieć* is a neologism that means ‘a net of words’.

We planned to automate part of the development effort, but we assumed that a *core* of about 7000 LUs would be constructed completely manually, as in the end it was. We did not take any monolingual dictionary as a starting point. Instead, we decided to start with a large corpus – the IPI PAN Corpus (Przepiórkowski, 2004) [IPIC], the largest available corpus of Polish, about 254 million tokens – and to extract a list of LUs for the core plWordNet directly from IPIC. The only criteria were part of speech and the frequency of basic morphological forms corresponding to particular LUs. The initially extracted list of over 10000 lemmas was manually filtered during preparation for the linguistic work. Section 2.4 discusses the work procedure and the drawbacks of a purely corpus-based approach. The nominal part of the core plWordNet was intended to cover the upper hypernymy levels, but it turned out that neither manual filtering of the initial frequency-based list nor subsequent extension of the list with LUs translated from the top levels of PWN ensured such coverage. During semi-automatic work – see Section 4.5 – we discovered many initially overlooked higher-level hypernyms.

It had been our intention to use the core plWordNet as a starting point for a form of bootstrapping. We assumed that the remainder of the initial plWordNet would be built semi-automatically, thus helping lower labour-intensity. Section 4.5 presents the *WordNet Weaver*, a software tool that combines several algorithms for the extraction of lexico-semantic relations. Section 4.5.4 discusses its largely positive effect on the linguists' performance. Most algorithms developed for the WordNet Weaver (Section 3) were evaluated (Section 3.3), and some also trained (Section 4.5.1), on the data acquired from the core plWordNet.

Before the start of the project, we ran preliminary experiments in the automatic extraction of synonyms from a large corpus. They led us to expect lower accuracy for more general and for less frequent LUs. It now turns out that the first guess was mostly inaccurate (Section 3.4) but the second was mostly true, except for manually constructed extraction patterns (Section 4.1).

In keeping with our long-term goal of developing a valuable lexical resource for Polish, we insisted all along on the trustworthiness of plWordNet. That is to say, we could not rely on fully automatic construction of the wordnet. The familiarity with previous work in this area left no doubt that manual correction of the extraction results would be indispensable. We revisit this issue in Section 4.5.4. Moreover, in the expectation of lower accuracy for more general LUs, we focused more on automated expansion of the core plWordNet than on the construction of some parts of the wordnet from scratch.

In the second phase of the project, we wanted to expand the core plWordNet semi-automatically with a relatively large number of new LUs and thus reach a size of no fewer than 15000 and no more that 25000 LUs. We expected that the manual correction of the automatically proposed plWordNet expansions would be selective control rather than extensive correction. We look at this assumption in Section 4.5.4.

According to our initial plans, an extraction algorithm should suggest both new synsets and instances of lexico-semantic relations. In the end, the WordNet Weaver generates only suggestions of *attachment points* (Section 4.5.3): synsets in which a given new LU can be included or to which it can be attached as a new hyponym/hypernym or even meronym. The accuracy of clustering-based methods of suggesting new synsets ended up too low for practical applications (Section 3.5). The use of support tools notwithstanding, we wanted to abide by the principle that the ultimate responsibility for every wordnet element rests with its authors in every phase of the wordnet development. It was tempting to speed up the development of our wordnet at the cost of slightly lower accuracy, but we are convinced that a smaller wordnet with excellent accuracy is more useful in applications than a larger but less reliable resource.

Despite the limited funds, we fully expected to build a wordnet of a size comparable to several much better established European wordnets. The introduction of the automated methods in the second phase of the project was meant to reduce the linguistic workload considerably[8]. Section 4.5.4 reports on the extent to which this succeeded.

There are many methods of extracting lexico-semantic relations from corpora. We present an overview and a detailed discussion of selected methods throughout Chapters 3 and 4. They can be roughly divided into two main groups of methods, basedon distribution (Chapter 3) and on patterns (Chapter 4). The former can achieve a relatively good accuracy in extracting instances of hypernymy – pairs of LUs – but very rarely of other relations such as synonymy, meronymy or antonymy; the recall is low. Distributional methods achieve good recall, because they can generate a description for any pair of LUs, but their accuracy is quite low: they do not distinguish between different lexico-semantic relations and produce a vague measure of semantic relatedness.

A well-known weakness of distributional methods is in distinguishing different LUs for the given *lemma*. Henceforth, we will understand lemma to be a *basic morphological word form* that represents the occurrences of one or a few particular LUs in language expressions. A lemma is monosemous if it represents one LU, and polysemous otherwise. The basic morphological word form, or *base form*, is a word form or language expression with conventional values of grammatical categories, such as the nominative case and singular number for nouns. A base form represents a set of word forms with the same meaning and different values of grammatical categories. We decided to operate on lemmas during the extraction of relation instances, because the number of different word forms is very high in the strongly inflected Polish language. Lemmatisation, or the mapping of word forms to lemmas, must be done automatically

[8] That is why we have allotted the funds approximately in the proportion 1:2 to manual work and to the software design and development work.

for large corpora; some error ratio is inevitable. We will discus corpus preprocessing in Section 3.4.3.

That is why we assumed from the start that it will be necessary to construct hybrid solutions: combine several methods, at least one following the pattern-based paradigm and one based on Distributional Semantics, see Section 3.2. We had been sceptical – justifiably, as Section 3.5 shows – about the possibility of recognising different LUs represented by a lemma on the basis of semantic clustering of lemmas. We therefore also planned to develop sense extraction for lemmas by clustering documents or at least longer segments that include occurrences of particular lemmas. We assumed that polysemous lemmas would occur in several documents. This part of our initial plans was the least successful (Section 3.5), but the other hybrid methods, when combined in the WordNet Weaver, achieved a level sufficient for practical application in the linguists' work.

1.3 Early Decisions

1.3.1 Models for wordnet development

PWN began as a psychological experiment and gradually morphed into a large ongoing lexical resource project. We naturally tried to explore the accumulated effects of long-term work on PWN, but the EWN project (Vossen, 2002) also attracted our attention. EWN aimed to develop a family of aligned wordnets (Section 1.1.4), and the scale of the enterprise required careful design. The EWN team also had an opportunity to analyse the previous PWN experience. All of this made the EWN project an important reference point for us.

There is a fundamental difference between the EWN and plWordNet projects: the former was oriented toward the development of aligned wordnets, while the present stage of plWordNet construction focusses on the appropriate description of Polish. We leave the question of mapping onto other wordnets for the upcoming continuation of the present plWordNet project[9]. The question of the appropriate sense-relating two-way mapping of wordnets for pairs of languages influenced how EWN constructed the wordnets. The solution was to link by expressing, in particular wordnets, the same lexicalised concepts from a shared set using the Inter-Lingual Index (Section 1.1.4). Besides this strategy, which somehow imposed seeking out lexicalisation of the same concepts in each language considered, two basic models of wordnet development have been worked out in EWN (Vossen, 2002, pp. 52):

[9] The budget of our project was too limited to investigate the problems of mapping (or, regrettably, to write glosses).

Merge Model: the selection is done in a local resource and the synsets and their language-internal relations are first developed separately, after which the equivalence relations to WordNet 1.5 are generated.

Expand Model: the selection is done in WordNet 1.5 and the WordNet 1.5 synsets are translated (using bilingual dictionaries) into equivalent synsets in the other language. The wordnet relations are taken over and where necessary adapted to EuroWordNet. Possibly, monolingual resources are used to verify the wordnet relations imposed on non-English synsets.

It has been observed that the expand model can lead to a wordnet biased by WordNet 1.5. For many languages, however, either no electronic monolingual resources – extended monolingual dictionaries or thesauri – are available, or existing resources are small, often with limited information in their entries. There have been suggestions that for such languages the expand model can work well in wordnet development. In the scope of EWN, the expand model was adopted for the Spanish and French wordnets. Later several other wordnet development projects also adopted it, including the Croatian WordNet (Raffaelli et al., 2008) and Hungarian WordNet (Miháltz et al., 2008).

A wordnet constructed following the merge model should provide a description of lexico-semantic relations closer to the spirit of the given language, in that it is less influenced by the design decisions in a wordnet for another language (probably English), often of a significantly different type. The merge model, however, requires rich resources at the outset, for example, a monolingual dictionary with senses identified, detailed definitions, thematic codes for senses and some semantic structuring. Such resources are created for humans readers, so to construct a wordnet from them is more than merely a matter of copying[10] – see (Pedersen and Nimb, 2008) for the use of resources in the DanNet project. The difference is also clear when one compares PWN and LDCE (Bullon et al., 2003), or plWordNet and (Dubisz, 2004).

1.3.2 Why we chose the merge approach

No electronic dictionary on which we could base the construction of Polish wordnet was available[11]. In addition, we did not want to consider indiscriminate mapping of PWN, and we dismissed the idea of translating it into Polish. In effect, we decided to build plWordNet from scratch. On the other hand, we wanted to keep plWordNet

[10] If a dictionary contains rich information structured in a way that facilitates NLP, we face another question: is the wordnet the best way of describing lexical semantics for NLP? We have no experience to answer such a question, because the Polish language, unfortunately, is not blessed with such abundance.

[11] The existing Polish electronic dictionaries, for example (Dubisz, 2004) or (PWN, 2007), are not freely available for research, and in any event their structure makes their usefulness limited.

compatible with PWN and at the same time have it appropriately reflect the relations in the Polish lexical system. We tried to adopt the PWN and EWN relation structure as much as possible, but 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, p. 86).

The Czech team noticed problems with the translation of equivalents and the corresponding gaps with regard to English. They observed two cases where it was not possible to find a synonyms (or even a near-synonym[12]). The Czech synsets had no lexical equivalents in English because of the difference in lexicalisations and conceptualization, or because of the typological differences between those two languages. There are, for example, no phenomena in English to correspond to the Czech verb aspect, reflexive verbs or rich word formation. It is widely assumed 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 (Wierzbicka, 2000, p. 193). We did try to translate the higher hypernymy levels of PWN (Section 2.4), only to discover four serious problems.

1. Some entries from the higher hypernymy levels of PWN (also called "strings" there (Miller et al., 2007)) can hardly be considered to denote frequent, basic or most general concepts in Polish; examples include *skin flick* 'film pornograficzny'[13], *party favour* 'pamiątka z przyjęcia', *butt end* 'grubszy koniec', *end, remainder, remnant, oddment* 'resztka materiału', *apple jelly* 'galaretka jabłkowa'.

2. PWN glosses are not always precise enough to let us find the Polish equivalent, or there may be no Polish equivalent at all; examples of untranslatable synsets include {`incolubrid snake, colubrid, elapid, elapid snake`}, {`communicator, acquirer`}.

3. Translating PWN would create nodes in the hyponymy/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', *transparent substance, translucent substance* 'materiał półprzezroczysty', *states' rights* 'prawa stanowe'.

[12] The term *synonym* and therefore also the term *near-synonym* are quite vague. Synonyms are discussed in Section 2.1; a near-synonym can be defined as a LU substitutable in a context, but strongly marked by its expressiveness, genre and so on, for example, *a girl* and *a chick*.

[13] All examples in this section are given Polish equivalents as 'glosses'.

4. On the other hand, some Polish LUs have no English lexical equivalents[14]: *brodacz* ‘bearded man’, *doczytać* ‘to read to the end’, *płaskodenny* ‘with a flat bottom’, *walizeczka* ‘small suitcase’. We decided, therefore, to describe the lexicalisation and conceptualization in Polish as accurately as possible. We believe that it is much more interesting to compare two wordnets that reflect the real nature of two natural languages than to create a hybrid, which in fact would be just an English wordnet mechanically translated into Polish.

In addition to opting for the merge model, we made several more detailed decisions for plWordNet.

- Synonymy, antonymy, hypernymy and meronymy hold between LUs of the same morphosyntactic class – nouns, adjectives or verbs; this is the basic assumption in PWN and EWN.

- Relations are divided into two subclasses: those linking synsets and those linking LUs; this was the most erroneous decision of all (Sections 2.1, 2.4), although it strictly followed the practice in previous wordnets. We backed away for the purposes of linguistic work before it even started, but the decision affected the application that supported wordnet construction (Section 2.4).

- Meronymy is divided into six subclasses, following EWN.

- Due to the strong potential of Polish lexical derivation, some relations were added or redefined to cover the most frequent or regular phenomena (Sections 2.2.4 and 2.2).

- Because we cannot add glosses to the entries in the databases, we decided to define an entry as a certain graphical string with a net of relations representing meaning; in consequence polysemy increased (Chapter 5.2).

[14] Naturally it is possible to employ a syntactic structure to express in another language approximately the same sense as the lexicalised term has, it just would not be a lexeme and not even an idiom.

Chapter 2

Building a Wordnet Core

2.1 The Synset

Synonymy plays a central role in the Princeton WordNet [PWN]. It is often referred to as a "basic semantic relation" in PWN – see for example (Miller, 1998, p. 23). The basic building block of PWN is a *synset*, presented as a "set of synonyms" (*ibid.*) or "a set of words with the same part of speech that can be inter-changed in a certain context" (Vossen, 2002, p. 5). The synset is also meant to be a vehicle for a *lexicalised concept* (Miller et al., 1993). It is sometimes defined as a set of lexical units which refer to the same lexicalised concept – and lexicalised concepts are presented as objects described, via synsets, by "conceptual-semantic relations" (Fellbaum, 1998a, p. 210).

It is a very problematic exercise to try and define synsets by means of lexicalised concepts: the latter notion is singularly vague. Besides, a whiff of circularity hangs over the whole terminology. A *concept* better be defined without referring to linguistic terms. This can, in principle, be done by applying methods in formal semantics, but it is hard to do it with a substantial portion of the vocabulary. A big advantage of a wordnet is that we can construct it without describing lexical meanings formally. It is more practical to go the other way around: from a synset to a lexicalised concept.

Since a synset is commonly defined through synonymy, let us look at that notion. There are two styles of synonymy definition (Derwojedowa et al., 2008): refer to mutual substitutability in a context, or derive synonymy from the hypernymy relation. In the former style, two words A and B are synonyms if, in a given context, A can be substituted for B and B for A without affecting the overall meaning. This type of synonymy often underlies the definition of synsets – see Vossen (2002, p. 5) cited above. The difficulty is with the notion of context. A context is typically defined by an example sentence, and one considers its meaning with and without the substitution. There is, however, linguistic evidence that strict synonymy does not exists (Bloomfield, 1933, pp. 145), (Hockett, 1964, Sec. 15.1), (Lyons, 1989, Section 9.4), (Apresjan, 2000, pp. 207) or (Edmonds and Hirst, 2002), so any substitution changes the meaning *somewhat*. An acceptable range of changes must therefore be accounted for in any synonymy definition – via some extralinguistic properties – or a reference to the linguist's intuition is required: how unimportant the change which the substitution introduces really is. That is a rather vaguely delineated task.

The second style of synonymy definitions is based on mutual hypernymy (Lyons, 1989, Section 9.4). If A is a synonym of B, then "A is a kind of B" and "B is a kind of A". For example (Derwojedowa et al., 2008), *ascending* is a kind of *going up* and the other way around, and so are *animal* and *beast*. Though *girl* is a kind of a *woman*, however, not all women are girls. Synonymy test can be assisted a substitution test of the kind we present in Appendix A (actually applied in plWordNet). We believe that a definition based on mutual hypernymy allows more subtle and less arbitrary discrimination of synonyms (or near-synonyms): the question asked in the test requires a simple yes-no answer and does not enforce an evaluation of the change.

This summary of Sections 1.1.1 and 1.1.2, together with a brief overview of synonymy definitions, emphasises the main points to which every wordnet designer must refer. Usually the structure of a wordnet strictly follows the PWN assumptions. A wordnet, however, tends also to be treated – and used – as a useful language resource that describes lexical semantics. The organisational principles must be clear when it comes to the fundamental unit of description – the word (with the inevitable language-dependent differences of opinion on what constitutes a word). There is justified doubt whether the synonymy relation and relations between concepts are a basis precise enough to be the underpinning of a wordnet.

The ontological and psychological status of a *concept* is not clear, nor is the relation between the concept, the word and the world. It is well known that expressions can have the same referent but different meaning, so they cannot be considered synonymous, as in Frege's famous pair "the morning star" and "the evening star".

Some word forms can have the same designative meaning but different expressive meaning. For example, *ręka*, *łapa*, *graba* and *grabula* all mean 'hand'. Only *ręka* is neutral, and can be described as a meronym of *ciało* 'body' and a holonym of *dłoń* 'palm', *ramię* 'shoulder", *przedramię* 'forearm'.[1] On the other hand, *graba* 'mitt' does not have such meronyms or holonyms.

Some word forms freely replaceable in many contexts are *not* synonymous: consider *I was bit by a bulldog/dog*. In fact, *bulldog* is a hyponym of *dog*. The word *mak* 'poppy' can be accurately described as denoting a flower, weed or herb, but it does not mean that a flower, a weed and a herb are synonyms (Derwojedowa et al., 2008).

In PWN, semantic relations (except antonymy and derivational relations) hold between synsets – that is to say, between lexicalised concepts – rather than between word forms (Fellbaum, 1998a, p. 210). The lexicalised concept, however, is characterised only as an unspecified, abstract semantic object which represents the part of the meaning of synset members that is common to all of them (Miller et al., 1993). A relation between concepts, therefore, is a relation defined in the space of abstract objects; its association with the lexical meaning relations is not direct or obvious. Without a pre-

[1] In the colloquial usage, *ręka* refers to the shoulder, arm, forearm and hand together.

cise description of lexicalised concepts it is hard to formulate an evaluation procedure for testing whether a given pair of concepts is an instance of the given relation. The tests used in EWN Vossen (2002) refer to pairs of words, are defined in the space of word pairs and clearly originate from the well-known lexicographical practice.

In plWordNet, all lexico-semantic relations hold between lexical units which are the basic building blocks of the wordnet.

A *lexical unit* [LU] is a word in a broad sense: it may be an idiom or even a collocation, but not a productive syntactic structure (Derwojedowa et al., 2008). It is a string that has its morphosyntactic characteristics and a meaning as a whole. As a result, substrings within a LU have no meaning or inflection of their own, so they can be treated just as morphemes are treated inside a morphological structure (Derwojedowa and Rudolf, 2003). In other words, a LU is syntactically non-compositional (is a terminal), but not necessarily semantically non-compositional. A LU is a basic morphological word form (see the definition on page 17, Section 1.2) *and* its meaning. There is, for example, *zamek 1* ‘castle’ and *zamek 2* ‘lock’. The basic morphological word form understood in this technical way will henceforth be referred to as a *lemma*. There are several methodological reasons why we decided to follow the traditional lexicographic approach – see also (Derwojedowa et al., 2008).

We treat synonymy more restrictively than PWN: LUs can be considered synonymous if they have the same hypernym, holonym or meronym. For example, *chaber*, *bławatek* and *modrak* are synonyms, because they all denote the same object ‘cornflower’ and share all lexico-semantic relations. On the other hand, *warzywo* ‘vegetable’ and *włoszczyzna* ‘vegetable bundle for soup’ cannot be consider synonymous: *włoszczyzna* consists of several very specific vegetables (each of them would be a meronym of *warzywo*).

We put a given LU into a *synset* because of all lexico-semantic relations of this LU with other units in the network (Derwojedowa et al., 2008). For example, *mak 1* ‘poppy *Papaver*’ is a hyponym of *roślina* ‘plant’ and a holonym of *makówka* ‘poppy head’; *mak 2* ‘poppy seed’ is a hyponym of *nasienie* ‘seed’ and a meronym of *makowiec* ‘poppy-seed cake’. A wordnet is a network of LUs connected by lexico-semantic relations. LUs with the same pattern of relation instances (such as linking to the same LUs via central lexico-semantic relations, notably hyperhymy/hyponymy or holonymy/meronymy) are grouped into synsets. A synset is therefore a “short cut” for two or more LUs which share a set of relations. Such view of the basic building blocks affects the structure of plWordNet considerably: synsets tend to be quite small, the semantic similarity of synset members is strict and many (especially nominal) synsets have just one element.

2.2 The Lexico-semantic Relations

A set of lexico-semantic relations that underlie a wordnet is its most distinguishing design consideration. While languages with different typology require subtly different sets, many relations carry well across types. For clear portability reasons, we decided to stay as close as possible to the PWN set of relations, and to include a few from the EuroWordNet [EWN] project (Vossen, 2002). The current version of plWordNet supports the following relations (the last two come from EWN):

- synonymy,
- antonymy,
- conversion,
- hypernymy/hyponymy,
- troponymy,
- holonymy/meronymy,
- relatedness,
- pertainymy,
- fuzzynymy.

We have kept the division of LUs into grammatical classes (parts of speech, as in PWN): nouns, verbs and adjectives. Relations other than *relatedness* and *pertainymy* connect LUs in the same class. Some relations are symmetrical (for example, if A is an antonym of B, then B is an antonym of A) or are mutual inverses (for example, a hyponymy pair is always the inverse of the corresponding hypernymy pair), while others are not (for example, holonymy: *a spoke* is part of *a wheel*, but not every wheel has spokes). We refer to both these properties of semantic relations by the general term *reversibility*, and assign it the value "$+$" or "$-$". The value is "$-$" only for meronymy-holonymy pairs and troponymy-hypernymy pairs, the latter because plWordNet distinguishes troponymy from hyponymy – see Section 2.2.2.

Following EWN, we have defined substitution tests for each relation. The tests are meant to be a tool that illustrates the definition, facilitates identification of relation instances and promotes consistency of decisions among linguists. The tests are presented in Appendix A.

Similar to other wordnets, among them PWN and EWN, lexico-semantic relations are defined in two domains: LUs and synsets. Hyperhymy/hyponymy and holonymy/meronymy are defined in the domain of synsets: they are subsets of the Cartesian product of the set of synsets. The other relations are defined in the domain of LUs. In contrast with most wordnets, however, the synset relations are not conceived as relations between lexicalised concepts but originate directly from the corresponding linguistic relations which hold between members of the respective synsets. An instance of a synset relation is a kind of short cut which expresses the existence of instances of the corresponding linguistic relation.

2.2.1 Antonymy and conversion

We have a very wide definition of *antonyms*:

- typical "opposition" pairs such as *mądry* 'wise' ↔ *głupi* 'stupid';
- pairs of complementary concepts such as *siostra* 'sister' ↔ *brat* 'brother' or *homoseksualista* 'homosexual' ↔ *heteroseksualista* 'heterosexual';
- opposite orientations such as *północny* 'northern' ↔ *południowy* 'southern' or *przedni* 'frontal' ↔ *tylny* 'rear';
- culturally motivated juxtapositons such as *ciało* 'body' ↔ *dusza* 'soul'.

Some LUs, particularly nominal units, have more than one antonym. For example, *mowa* 'speech' is an antonym of *pismo* 'writing' but also of *milczenie* 'silence'. Even more interesting is the example of *spokój* 'calm' with several antonyms: *agresja* 'aggression', *gniew* 'anger', *lęk* 'anxiety', *niepokój* 'uneasiness', *szaleństwo* 'craziness' and *złość* 'fury'.

Antonymy links *strictly* a pair of LUs in plWordNet, so we only define it at the level of LUs. We have also often noted that a definition of antonym links forces a "splitting" of an LU, each version with different antonyms. This, in turn, resulted in additional synsets and a more fine-grained description of polysemous lemmas. Our flexible definition of synsets allows even the introduction of LUs which do not belong to any synset.

We keep antonymy (*good* ↔ *bad*) separate from *conversion* (*wife* ↔ *husband*), which is described as a separate relation specific to plWordNet. That is because, following Apresjan (2000, Section 6, pp. 242-265), we believe that conversion differs from synonymy and antonymy. For example, the verbs *kupić* 'buy' and *sprzedać* 'sell' describe the same situation (a commercial transaction), but they portray it from different points of views. The meaning of one LU logically arises from the meaning of the other: if X buys something from Y, Y sells it to X. The motivating (and very interesting) examples of conversion are the words *dziewczyna* 'girl' and *chłopak* 'boy'. In Polish – and similarly in English – a juxtaposition of these nouns means either *girl* ↔ *boy* or *girlfriend* ↔ *boyfriend*. That is to say, the relation should be either antonymy or conversion. It was therefore essential to create two LUs: *dziewczyna 1* ‘girl’ and *dziewczyna 2* ‘girlfriend’.

2.2.2 Hyponymy/hypernymy and troponymy

The central *hyponymy/hypernymy* relation shapes the hierarchical structure of the lexicon. It mandates the formation of long superclass-subclass paths. One small example will illustrate: *animal* → *dog* → *poodle* → *toy poodle*. The relation is prevalent among nouns, especially where it comes to representing natural types and role types. Let us offer a detailed analysis of one “family” of concepts. *Roślina* ‘plant’ has *organism* ‘organism’ as a hypernym, and several hyponyms: *krzew* ‘bush’, *drzewo* ‘tree’, *trawa* ‘grass’, *glon* ‘alga’, *alga* ‘alga’ (the last two are synonyms) and *roślina uprawna* ‘cultivated plant’. Most of these hyponyms have their own hyponyms. Thus, *roślina uprawna* ‘cultivated plant’ includes *zboże* ‘cereal’ and *warzywo* ‘vegetable’. The latter has hyponyms such as *por* ‘leek’, *kapusta* ‘cabbage’ and *ziemniak*, *kartofel*, *pyra*, *grul*, all four meaning ‘potato’. The same goes for *drzewo* ‘a tree’: its hyponyms include *drzewo iglaste* ‘conifer’ and *drzewo liściaste* ‘deciduous’, each with numerous hyponyms.

In plWordNet, hyponymy/hypernymy also holds among verbs. For example, *okazywać uczucia* ‘express one’s feelings’ has the following hyponyms: *wzruszyć się* ‘be moved’, *uśmiechnąć się* ‘smile’, *zabawiać się* ‘divert oneself’, *rechotać* ‘chortle’, *śmiać się* ‘laugh’, *tulić uszy* ‘back down (literally *fold one’s ears*)’, *ucieszyć się* ‘rejoice’, *wylać łzy* ‘shed tears’, *wyśmiewać się* ‘mock’, *złościć się* ‘be angry’, *zezłościć się* ‘get angry’ and *zdziwić się* ‘be surprised’.

Hyponymy among verbs in PWN is identified with the *troponymy* relation, and forms a symmetrical pair with hypernymy. Troponymy is “a manner relation”, with the following description in Fellbaum (1998b, p. 79):

To V_1 is to V_2 in some particular manner.

In EWN (Vossen, 2002), troponymy was replaced with hyponymy among verbs. We see a place for both relations. Verbs which describe the manner of action, such as *mówić* ‘speak’ ← *jąkać się* ‘stammer’, are linked by *troponymy*. Here are other examples of troponymy: *iść* ‘walk’ is linked in plWordNet with *kroczyć* ‘stride’ and *leźć* ‘trudge, shamble’; *przykrywać* ‘cover’ with *pokrywać* ‘coat’ and *okrywać* ‘wrap (in)’; and *brać* ‘take’ with *zabierać* ‘take away’. It must be noticed that the relation is not symmetrical. While ‘trudge, shamble’ can be paraphrased as ‘walk in certain way’, it would be wrong to describe *iść* in the same way with respect to *leźć*.

The majority of Polish troponyms are morphological derivatives created by a set of prefix morphemes from their hypernyms as their derivative bases (Derwojedowa and Zawisławska, 2007a). Troponymy and hypernymy are, then, defined in the domain of LUs, not synsets. In the literature (Lyons, 1989) one can find claims that hyponymy does not occur among verbal LUs, but only links gerunds derived from the verbs by regular derivation processes. Derwojedowa and Zawisławska (2007a) argue, however, that it is necessary to distinguish between the cases of meaning specialisation represented by verbal hyponymy and troponymy expressed by a derivational link.

The substitution test for verbal hypernymy/hyponymy, presented in Appendix A, page 187, refers to the semantic entailment of the hypernym by the hyponym, but also to the presence of the hypernymy/hyponymy relations between the respective gerunds. The substitution test for troponyms – page 188 – differs in two ways. First, we expect that the entailing sentence can be extended to a paraphrase which includes an additional modifier of manner. Second, the pair of respective gerunds – derived from a verbal hypernym and troponym – is not an instance of the nominal hypernymy/hyponymy.

A little unexpectedly, hyponymy/hypernymy is relatively widespread for adjectives (*karminowy* 'crimson' → *czerwony* 'red'), and particularly common among relational (desubstantival) adjectives. Examples of hyponymy/hypernymy can be found among qualitative adjectives as well: *mleczny* 'made of milk' → *spożywczy* 'alimentary', or *brunatny* 'russet, tawny' → *brązowy* 'brown'. An elegant example of a hyponymy/hypernymy tree for adjectives (the deepest thus far in plWordNet) is the synset *europejski* 'European'. It has the following hyponyms:

- *austriacki* 'Austrian',
- *litewski* 'Lithuanian',
- *niemiecki* 'German',
- *hiszpański* 'Spanish',
- *węgierski* 'Hungarian',
- *brytyjski* 'British' (with *angielski* 'English' as its hyponym),
- *francuski* 'French' (with one hyponym, *paryski* 'Parisian'),
- *skandynawski* 'Scandinavian' (with two hyponyms, *norweski* 'Norwegian' and *szwedzki* 'Swedish'),
- *włoski* 'Italian' (one hyponym: *rzymski* 'Roman'),
- *słowiański* 'Slavic' (with four hyponyms: *czeski* 'Czech', *ukraiński* 'Ukrainian', *rosyjski* 'Russian' (with *sowiecki* 'Soviet' as *its* hyponym), *polski* 'Polish').

Rather naturally for a Polish resource, there are additional details for the adjective *polski* ‘Polish’. It has three hyponyms:

- *śląski* ‘Silesian’,
- *mazowiecki* ‘Mazovian’ (with *warszawski* ‘from/of Warsaw’ as its hyponym),
- *małopolski* ‘from/of Lesser Poland’ (with two hyponyms, *krakowski* ‘Cracovian’ and *oświęcimski* ‘from/of Auschwitz’).

Sometimes lexical gaps occur in the hyponymy/hypernymy hierarchy. There are groups of LUs closely related as denoting kinds or forms of something, but there is no LU to denote their common hypernym (i.e. a LU existing in Polish). We fill such gaps with *artificial LUs*, following the practice in GermaNet (Hamp and Feldweg, 1997). An artificial unit is a syntactic construction, not lexicalised in Polish. For example, the noun LU *człowiek* ‘human’ dominates a lexico-semantic relation tree with more than 20 artificial units. They include:

- *człowiek ze względu na swoje zajęcie* ‘human with regard to occupation’;
- *człowiek ze względu na płeć* ‘human with regard to sex’;
- *człowiek ze względu na kwalifikacje* ‘human with regard to qualifications’ (the hyponyms include *amator* ‘amateur’ and *ekspert* ‘expert’);
- *człowiek ze względu na sytuacje materialną* ‘human with regard to financial condition’ (with the hyponyms *pan* ‘lord’, *pani* ‘lady’, *bogacz* ‘rich man’, *biedak* ‘poor man’ (with its hyponym *żebrak* ‘beggar’));
- *człowiek ze względu na swoje cechy* ‘human with regard to personal features’ – it is the root of a larger hyponymic cluster:
  - *człowiek oceniany pozytywnie albo negatywnie* ‘human perceived positively or negatively’,
  - *człowiek charakteryzujący się jakąś cechą* ‘human characterized by something’,
  - *człowiek ze względu na wiek* ‘human with regard to age’;
- *człowiek ze względu na relacje społeczne* ‘human with regard to social relationships’ (with *członek* ‘member’ and *członek rodziny* ‘family member’ as hyponyms).

Artificial units also appear among verbs. The largest hyponymy/hypernymy tree for verbs contains *wykonywać czynności prawne* ‘perform legal activities’ as the direct hypernym of the following hyponyms:

- *wypuścić* 'set free', *sądzić* 'judge', *sędziować* 'be a judge',
- *rozstrzygać* 'adjudicate' (with *rozpatrzeć* 'investigate' as its hyponym);
- *umorzyć* 'dismiss (a case)',
- *unieważnić* 'annul' (with *odwołać* 'revoke' as its hyponym);
- *upoważnić* 'authorise', *zatwierdzić* 'approve'
- *głosować* 'vote' (with the hyponyms *uchwalić* 'pass (legislation)', *wstrzymać się* 'abstain' and *wyłonić* 'select').

The synset *wykonywać czynności prawne* 'perform legal activities' has one more hyponym and in it the artificial unit *zareagować na złamanie prawa lub normy społecznej* 'react to a breach of law or social norm' with several hyponyms: *aresztować* 'arrest', *karać* 'punish', *ukarać* 'punish', *skazać* 'convict' and *wymierzyć* 'impose (a sentence)'. All this shows that adding the main hypernym with the artificial unit *wykonywać czynności prawne* was necessary for buiding the tree of relations and describing the links between the other verbs properly.

In plWordNet we have tried to avoid mixing naive, popular classes with scientific categories. For example, we have two LUs for *cukier* 'sugar'. One is *cukier 1* with an antonym *sól* 'salt', hypernym *przyprawa* 'seasoning' and hyponyms *cukier puder* 'icing sugar', *cukier kryształ* 'granulated sugar', *cukier waniliowy* 'vanilla sugar' and so on. The other LU is *cukier 2* 'sugar' with a synonym *węglowodan* 'carbohydrate', hypernym *związek* 'compound' and hyponyms *fruktoza* 'fructose', *glukoza* 'glucose', and so on.

2.2.3 Meronymy/holonymy

The meronymy/holonymy relation is present in plWordNet for some nouns. Meronymy is a semantically diverse relation, so we have adopted the idea of meronymy/holonymy subtypes from PWN (Fellbaum, 1998c) and EWN (Vossen, 2002). The list of subtypes comes from EWN (Derwojedowa et al., 2007). In all examples, the meronym is shown first; the first example is generic, the second appears in plWordNet.

1. *part*: *finger* → *hand*, {`egzemplarz, okaz`} 'specimen' → {`kolekcja, zbiór`} 'collection, set';
2. *portion*: *slice* → *bread*, {`tost, grzanka`} 'toast' → {`chleb`} 'bread';
3. *place*: *oasis* → *desert*, {`termin, data`}'deadline, date' → {`czas`}'time';

4. *element of a collection*: *tree* → *forest*, {`kula, nabój`} 'bullet' → {`amunicja`} 'ammunition';

5. *substance*: *rubber* → *welly*, {`dachówka`} 'tile' → {`dach, zadaszenie`} 'roof, roofing'.

The category of body parts illustrates deep meronymy/holonymy very well. For example, the meronyms of {`ciało`}'body' include {`ramię`} 'arm', {`głowa`} 'head', {`serce`} 'heart', but also {`krew`} 'blood' and {`tkanka`} 'tissue'. These, in turn, are holonyms for other synsets, for example {`głowa`} 'head' is a holonym for {`twarz`} 'face', which is a holonym for {`oko`} 'eye', which is a holonym for {`źrenica`} 'pupil' or {`tęczówka`} 'iris'.

Often more than one subtype of meronymy accounts for a given synset. For example, {`drzewo`} 'tree' has several *part* meronyms: {`korzeń`} 'root', {`gałąź`} 'branch', {`pień`} 'trunk', {`korona`} 'tree crown'. On the other hand, {`drzewo`} 'tree' is an *element of a collection* meronym of {`sad`} 'orchard' and {`las`} 'forest'.

2.2.4 Relatedness, pertainymy and Polish derivation

We have broadened the *relatedness* and *pertainymy* relations, the only morphological relations in PWN and EWN. PWN has been constructed for English, so only the properties of this language were considered. Slavic languages, and specifically Polish, differ from English in important ways, not anticipated in the PWN structure.

The first problem is aspect. It is a grammatical category specific for Slavic languages, exemplified by perfective-imperfective pairs *kupić* 'buy (once); have bought' - *kupować* 'buy (habitually)' or *napisać* 'write (once); have written' - *pisać* 'write (habitually)'. Aspect makes the description of verbs troublesome, because not all senses of a polysemous verb must have perfective-imperfective pairs or it need not be the same perfective pair. For example, the verb *czytać* 'read (text, habitually)' can be paired with *przeczytać* 'have read (text)', but the same verb *czytać* 'read someone's thoughts or read in someone's eyes' should be paired with *odczytać* 'have read (in eyes)' or *wyczytać* 'have read (in thoughts)'.

Polish rich morphology mandates many substantially different verbal derivatives. For example, the verb *czytać* 'read' has the following derivatives:

- *czytywać* 'read (repeatedly, several times)',

- *zaczytać* 'read so often as to wear out',

- *odczytać* 'decipher or read out',

- *doczytać* 'read to the end or to read more',

- *poczytać* 'read for some time',
- *sczytać* 'proofread',
- *wczytać* 'read in',
- *wyczytać* 'learn something while reading',
- *rozczytać* 'decipher illegible writing'.

Diminutive and augmentative forms are another productive type of derivation in Polish. For example, the noun *kot* 'cat' has diminutive derivatives *kotek*, *koteczek*, *kotuś*, *kociątko*, *kicia*, *kiciunia*, *kiciuś*. The adjective *biały* 'white' has diminutive derivatives *bieluchny*, *bielusieńki*, *bieluśki*, *bieluteńki* and *bielutki*. Feminine derivatives of the masculine basic morphological word forms are also a regular phenomenon. In Czech WordNet (Pala and Smrž, 2004) names like that are described by the relations {`x has male`} and {`x has female`} (Derwojedowa and Zawisławska, 2007b); it also existed in PWN in its original version.

Other derivatives include names of individuals with certain property, such as *rudzielec* 'red-head' (from *rudy* 'red'), *śpioch* 'sleepyhead' (from *spać* 'sleep'); names of tools, such as *gaśnica* 'fire extinguisher' (from *gasić* 'extinguish'), *ścierka* 'dishcloth' (from *ścierać* 'wipe'); names of places, such as *siłownia* 'gym' (from *siła* 'strength'), *jadalnia* 'dining room' (from *jadać* 'eat (habitually)'); names of youngsters, such *kociak* 'kitten' (from *kot* 'cat'); expressive names, such as *kobiecina* 'woman (condescending)' (from *kobieta* 'woman').

We divide all this morphological treasure into two relations: *relatedness* and *pertainymy*. The former is for more regular derivatives:

- "clean" aspectual pairs (verbs which differ only in the information whether the action was perfective),
- gerunds derived from verbs,
- abstract nouns derived from adjectives (such as *mądrość* 'wisdom' from *mądry* 'wise'),
- causative verbs,
- relational adjectives,
- participles.

Pertainymy is for less regular phenomena:

- names of features, places, countries and nationalities,

- names of youngsters,
- feminine/masculine names,
- augmentatives, diminutives and expressive names.

2.2.5 Fuzzynymy

From EWN, we also adopted the *fuzzynymy* relation. It is meant for pairs of LUs which are clearly connected semantically, but which the linguist cannot fit into the existing system of more sharply delineated relations. As for nouns, some fuzzynymy relations appear to be regular and repeatable. For example, fuzzynymy links nouns that describe employees and their workplace:

- *lekarz* ‘physician’ or *pielęgniarka* ‘nurse’ and *szpital* ‘hospital’,
- *kustosz* ‘curator’ and *muzeum* ‘museum’,
- *ksiądz* ‘priest’ and *kościół* ‘church’,
- *listonosz* ‘postman’ and *poczta* ‘post office’,
- *burmistrz* ‘mayor’ and *ratusz* ‘town hall’.

Not all pairs related by fuzzynymy are so radically different. We also linked up pairs which could as well be classified as derivatives, for example:

- *sędzia* ‘judge’ and *sąd* ‘court’,
- *ambasador* ‘ambassador’ and *ambasada* ‘embassy’,
- *ogrodnik* ‘gardener’ and *ogród* ‘garden’,
- *rolnik* ‘farmer’ and *rola* ‘farmland’.

Another apparent connection is between names of objects and places where these objects usually reside, or between activities and places where they occur. A few examples:

- *obraz* ‘picture’ or *rzeźba* ‘sculpture’ and *wystawa* ‘exhibition’,
- *roślina* ‘plant’ or *kwiat* ‘flower’ and *ogród* ‘garden’ ,
- *spacer* ‘walk’ and *park* ‘park’,
- *uczyć się* ‘learn’ and *szkoła* ‘school’.

It turns out that fuzzynymy links LUs from different semantic categories and even LUs in different parts of speech. It links, for example, a verbal name of activity and a nominal name of a person: *odebrać* ‘receive’ ↔ *odbiorca* ‘recipient’, *pobierać* ‘collect’ ↔ *poborca* ‘collector’, *wytworzyć* ‘produce’ ↔ *wytwórca* ‘producer’; a verbal name of activity and its effect: *dorabiać się* ‘become wealthy’ ↔ *dorobek* ‘wealth’, *szukać* ‘search’ ↔ *odnaleźć* ‘find’; or a verbal name of activity and a nominal name of an object connected with this activity: *wychodzić* ‘exit’ ↔ *wyjście* ‘exit’ or *składować* ‘store’ ↔ *skład* ‘storehouse’.

2.3 Difficult Cases

We will conclude the detailed considerations of the lexical issues in plWordNet with a short list of serious problems that arose during the construction of the wordnet. It was, for example, difficult to categorise some words precisely and to disambiguate LUs. An unusually polysemous noun *ojciec* ‘father’ means a *parent*, a *monk* or an *author*. Relations between this noun and other LUs depend on the meaning of the word. The solution was to create three different LUs belonging to three different synsets with different sets of synonyms, hyponyms and hypernyms. Thus, the hypernym of *ojciec 1* ‘father’ is *rodzic* ‘parent’ and its holonym is *rodzice* ‘parents’; *ojciec 2* ‘father’ has *zakonnik* ‘monk’ as its hypernym; *ojciec 3* ‘founding father’ appears in the same synset as *autor* ‘author’, *twórca* ‘creator’ and *pomysłodawca* ‘originator of an idea’, and has *reżyser* ‘director’ and *projektant* ‘designer’ as its hyponyms.

Another example is the verb *zdawać*. It means either ‘hand over’ or ‘pass (exams)’. In plWordNet, *zdawać 1* ‘to turn over’ has synonyms *oddawać* ‘give back’ and *przekazywać* ‘hand over’, while *zdawać 2* ‘to pass’ is synonymous with *dostawać się* ‘be accepted’ (at a university). The adjective *ambitny* (which means either ‘ambitious’, ‘demanding’ or ‘thought-provoking’) is a similar case: *ambitny 1* ‘ambitious, aspiring’, from which the noun *ambicja* ‘ambition’ is derived, refers to people and is not connected with any other LU; *ambitny 2* ‘intelectually stimulating, innovative’ has *tandetny* ‘trashy’ and *komercyjny* ‘commercial’ as its antonyms; finally, *ambitny 3* ‘challenging’ has only one antonym *nieambitny* ‘undemanding’.

It is even more interesting when the linguist has to find the difference between metaphorical and literal meaning of a word. Again, it was necessary to create two LUs. That is why, for example, there are two nouns *policzek* in plWordNet: *policzek 1* ‘cheek’ is a holonym of *twarz* ‘face’; *policzek 2* ‘slap in the face’ has synonyms *obelga* ‘insult’ and *zniewaga* ‘affront’, among others.

2.4 The First 7000 Lexical Units

The plWordNet project depends crucially on an initial manually built small network. It has been our firm belief from the start that semi-automated construction of a wordnet from the ground up requires a completely trustworthy core to achieve acceptable accuracy. We assumed that automated methods will not perform well for more general LUs, and will be unable to extract the basic structure of the future plWordNet. The first assumption turned out to be too pessimistic. Measures of Semantic Relatedness (Section 3.4) produced results of lower accuracy only for some more general LUs, while for others the results were quite correct or even good. The second assumption, however, has been borne out by the experiments with state-of-the-art clustering algorithms for the extraction of synsets and possibly a hypernymy structure – see Section 3.5.

We planned a fully manual construction[2] of *core plWordNet* with approximately 7000 LUs. Those would be LUs with a general meaning, such as nouns located in the upper part of the hypernymy structure, including LUs which represent concepts such as *rzecz* 'thing' or *substancja* 'substance'. We had settled upon not translating any existing wordnet, and no monolingual dictionary in an electronic form was available to be leveraged as a source of the plWordNet structure, especially the hypernymy structure. That is why we initially decided to rely only on a large enough corpus. The best choice for Polish was the IPI PAN Corpus [IPIC] (Przepiórkowski, 2004) – the largest available corpus of Polish at the time when the project began. IPIC, designed as a corpus of general Polish, consists of about 254 million tokens and contains a range of genres, including literature, poetry, newspapers, scientific texts, legal texts and stenographic parliamentary records. It is not balanced: the last category dominates (Przepiórkowski, 2006).

We first extracted 10000 most frequent one-word lemmas[3] in IPIC 1.0, each tagged with a *grammatical class*[4] (in the technical sense, see page 17, Section 1.2). We collected more lemmas than the planned size of the core plWordNet, because we expected the list to shrink after manual revision. We divided them manually into 45 general semantic domains (26 nominal, 15 verbal, 4 adjectival) corresponding to the domains that label source files of Princeton WordNet [PWN] 1.5.

Simultaneously with the grouping, the linguists filtered out typos and rare lemmas whose high frequency was an artefact of errors in morphosyntactic tagging of IPIC 1.0. For example, the verb *maić* '≈ adorn with verdure' normally occurs very rarely in rather old fashioned constructions (the use of any finite form is hard to imagine). The

[2] There was logistical software support for the process, but all lexicographic decision were to be made by linguists.

[3] A method of extracting two-word lemmas was developed later.

[4] In the IPIC tagset, word forms are divided into 32 grammatical classes, a division more-fine grained than the traditional parts of speech.

morphosyntactic tagger used in IPIC 1.0 misinterpreted the following two situations:

- the word form *maj*, which represents two lemmas: $May_{case=nom}$, but also the imperative form of *maić*,
- the word form *mają*: $to\ have_{num=sg,per=3rd}$ – "non-past form" (Przepiórkowski, 2006), but also the non-past, singular, third-person form of *maić*.

The tagger only recognised the latter, causing a completely wrong high frequency of the verb *maić*. Such lemmas were not included in the final list. Proper names, pronouns, numerals and foreign words were excluded from the list, too. Another problem was a bias introduced by the lack of balance in IPIC 1.0. Many lemmas appeared to be unexpectedly frequent while others did not occur at the analysed top positions. For example, there were almost no animal names, but *lis* 'fox' was excessively frequent: it is the surname of a well-known journalist.

Lemmas were acquired as separate tokens, so we had to attach the reflexive marker *się* to some verbs. A few examples: *dziać* had to be "reconstructed" as *dziać się* 'to happen'; *oglądać* 'watch' and *oglądać się* 'look back' were both necessary. We had to reconstruct some multiword lemmas (representing fixed multiword expressions) which were separated into several tokens or from which only one constituent token was present on the list – for example, *piłka nożna* 'football'.

Putting LUs in general domains gave us a kind of initial sense disambiguation. Even inside a domain, more disambiguation was occasionally required. For example, *białko* denotes 'egg white' or 'protein'.

The top-level hierarchy in PWN 1.5 consists of 26 nominal, 15 verbal and four adjectival domains. We decided to extend the adjectival domains by adding *gradual* and *deverbal* adjectives to the original two domains of *relational* and *descriptive* adjectives. The first of the new domains contains adjectives which signal the intensity of a feature (for example, *maluśki* 'tiny'); the second domain contains participles. We found the top-level hierarchy not perfectly suited to Polish, but in the end the division turned out to be unnecessary. The domains only helped distribute work among linguists. They were never meant to be a tool of semantic description. Every linguist was given one domain at a time to work on.

After the initial LU list has been established, a group of linguists – with two experienced coordinators – constructed the first collection of synsets. The synsets, understood quite broadly, grouped closely related LUs. We needed that first version fast to help develop automatic methods (to base evaluation on, see Section 3.3) and to make some applications of the wordnet possible soon. This, however, has turned out to be a wrong decision. It influenced negatively the subsequent steps of plWordNet development. Constructed in a few months, the initial set of broad synsets became a "fact on the ground" which made it quite hard for the coordinators to introduce the hypernymy structure. Hypernymy among broadly conceived synset had not actually been a primary concern: various other dependencies had been introduced instead. It is our experience that it would have been much better to build simultaneously the structure of all wordnet relations.

At this stage, further disambiguation and correction of the original list of LUs was performed. The linguists worked out synsets wide enough for some hypernyms and quasi-synonyms to be listed as synonyms. The substitution test – the possibility of using two words in the same context – is not always precise enough to help distinguish these relations. Also, expressive and vulgar vocabulary was relocated to hyponymic synsets, obsolete vocabulary – removed.

The team of linguists was located in different cities, so we needed a system to support distributed work. A support system should not only enable flexible access and keep the integrity of the database, but, we assumed, also protect against inconsistencies and facilitate some management of the work, including options for reporting errors and tracing corrections. Both assumption have been heavily revised by practice, as it will be discussed shortly.

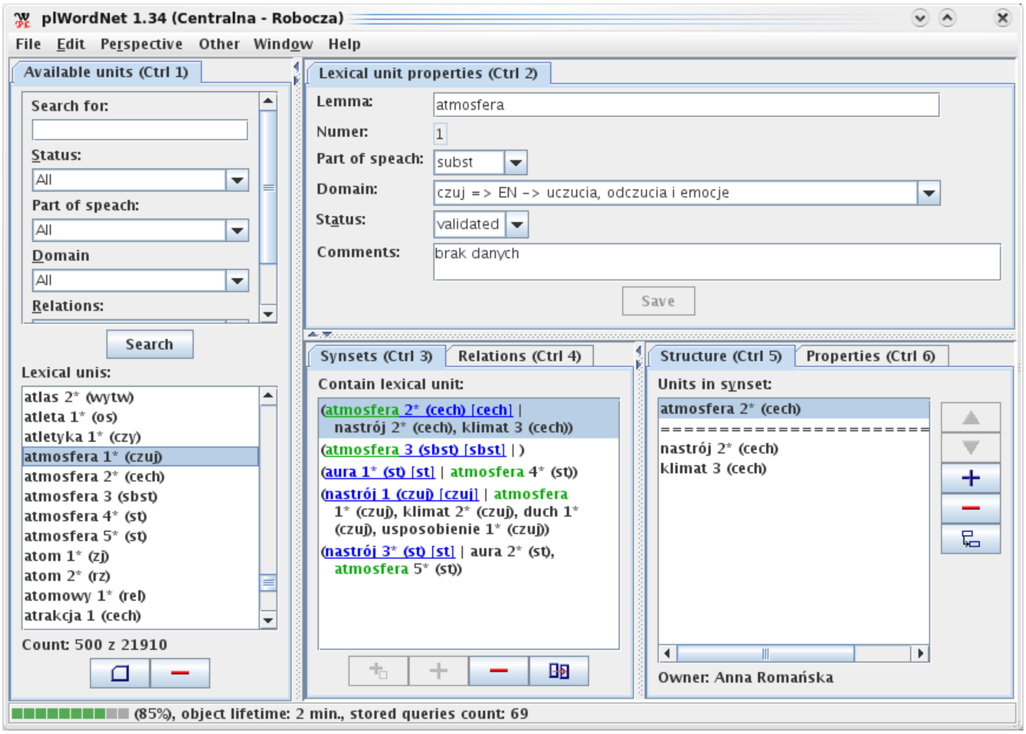
The story of wordnet editors begins with Grinder (Tengi, 1998), a software tool that checked the PWN source files and converted them into the lexical database. Linguists had to edit the source files. Syntactic and structural errors, such as pointers to nonexistent database elements, were identified only during compilation.

The EWN project (Vossen, 2002) constructed Polaris, an editor, and Periscope, a graphical database viewer. Both were commercial tools, tightly coupled with certain properties of the EWN database structure. The limitations of Polaris prompted the implementation of a new tool, VisDic, for the Czech WordNet project (Horák and Smrž, 2004). In VisDic, relation definitions are still written in text windows, but an XML format is used and some immediate browsing is possible in the tool, for example bi-directional browsing of graphs of semantic relations. VisDic is available for research.

VisDic was a monolithic application that worked directly on XML files. DEBVisDic (Horák et al., 2006), is a lexical database editor that reimplements and extends the functionality of VisDic. It is based on the client-server architecture and an XML database server. Both tools are oriented toward editing a wordnet synchronized with wordnets for other languages by the Interlingua Index (Vossen, 2002). That complicates their basic structure and user interface. Those characteristic features went beyond our needs, and anyhow DEBVisDic was not known to be available yet at the start of the plWordNet project. We decided, therefore, to build our own wordnet editor, *plWordNetApp* [plWNApp] (Piasecki and Koczan, 2007).

The plWordNetApp user interface was intended to support the division of work on plWordNet construction into the steps originating from the assumed plan of the whole process. The DEBVisDic tool is much more general: we designed plWNApp screens "minimalistically", for a set of particular tasks and users. As a consequence – an old truth – any error made during the planning of the tasks immediately decreased the final usability of the whole plWNApp in more than one way.

The Graphical User Interface [GUI] in plWNApp lets the linguists avoid the use of an artificial language for the description of semantic relations, starting with introduction of a new LU and its description. This improves on the practice in PWN and GermaNet (Fellbaum, 1998c, Hamp and Feldweg, 1997). All browsing and editing decisions are made via GUI screen controls and transparently recorded in the server or local database (depending on the selected mode). This tight coupling of GUI with the steps of the core plWordNet construction has influenced the basic division of the user interface into several main parts, called *perspectives*[5]. The most characteristic of them are the *LU perspective* and the *synset perspective*.



Search for:

Status: All

Part of speach: All

Domain: All

Relations:

Lexical unis:

atlas 2* (wytw)
atleta 1* (os)
atletyka 1* (czy)
atmosfera 1* (czuj)
atmosfera 2* (cech)
atmosfera 3 (sbst)
atmosfera 4* (st)
atmosfera 5* (st)
atom 1* (zj)
atom 2* (rz)
atomowy 1* (rel)
atrakcja 1 (cech)

Count: 500 z 21910

Lemma: atmosfera

Numer: 1

Part of speach: subst

Domain: czuj => EN -> uczucia, odczucia i emocje

Status: validated

Comments: brak danych

Contain lexical unit:

(atmosfera 2* (cech) [cech] | nastrój 2* (cech), klimat 3 (cech))
(atmosfera 3 (sbst) [sbst] |)
(aura 1* (st) [st] | atmosfera 4* (st))
(nastrój 1 (czuj) [czuj] | atmosfera 1* (czuj), klimat 2* (czuj), duch 1* (czuj), usposobienie 1* (czuj))
(nastrój 3* (st) [st] | aura 2* (st), atmosfera 5* (st))

Units in synset:

atmosfera 2* (cech)
========================
nastrój 2* (cech)
klimat 3 (cech)

Owner: Anna Romańska

(85%), object lifetime: 2 min., stored queries count: 69

Figure 2.1: The LU perspective

The LU perspective (Figure 2.1) was meant to support the grouping of LUs into synsets. The first version of plWNApp in 2005 had only this screen implemented. The list of LUs present in the system (the left panel) can be filtered according to

[5] This technical term has turned out to be infelicitous for users who are linguists.

several criteria, for example a selected domain. To facilitate search, each synset is also automatically assigned a domain according to its first LU. Once a LU has been selected, its relevant properties appear in the upper right panel. A new LU can be added (a button below the LU list) in the LU perspective – as well as in some other parts of the plWNApp GUI.

The description of a LU has the following elements: the *name* (lemma – plus an automatically generated unique sense number, *part of speech*, the *domain* (for organisational purposes), the linguist's *comment*[6], the *status* (the stage of processing such as for example "*completed by a linguist but not yet checked by a coordinator*"), and the *origin* (was the LU in the basic structure of the core plWordNet, or has it been introduced into some synset by a linguist?).

The description of a synset has the following elements: the *set of LUs*, the processing *status*, the linguist's *comment*, and the *artificial/standard flag* (it marks artificial LUs, see Section 2.2.2). Each LU and each synset have its unique automatically assigned identifier; plWNApp also checks whether an LU belongs to only one synset.

All synsets that include a given LU are shown in the tabbed panel below the LU property panel, see Figure 2.1 (the domains of the synsets are presented in blue). The editing of the selected synset is possible in the tabbed panel to the right of it – the second, hidden tab pane contains synset properties. We assumed that a selected LU is first assigned to an existing or a newly created synset, and next the synset is edited.

In the hidden tab pane of the synset list panel, one can browse and edit a list of lexico-semantic relations of the selected LU (between pairs of LUs, for example, antonymy or derivational relations).

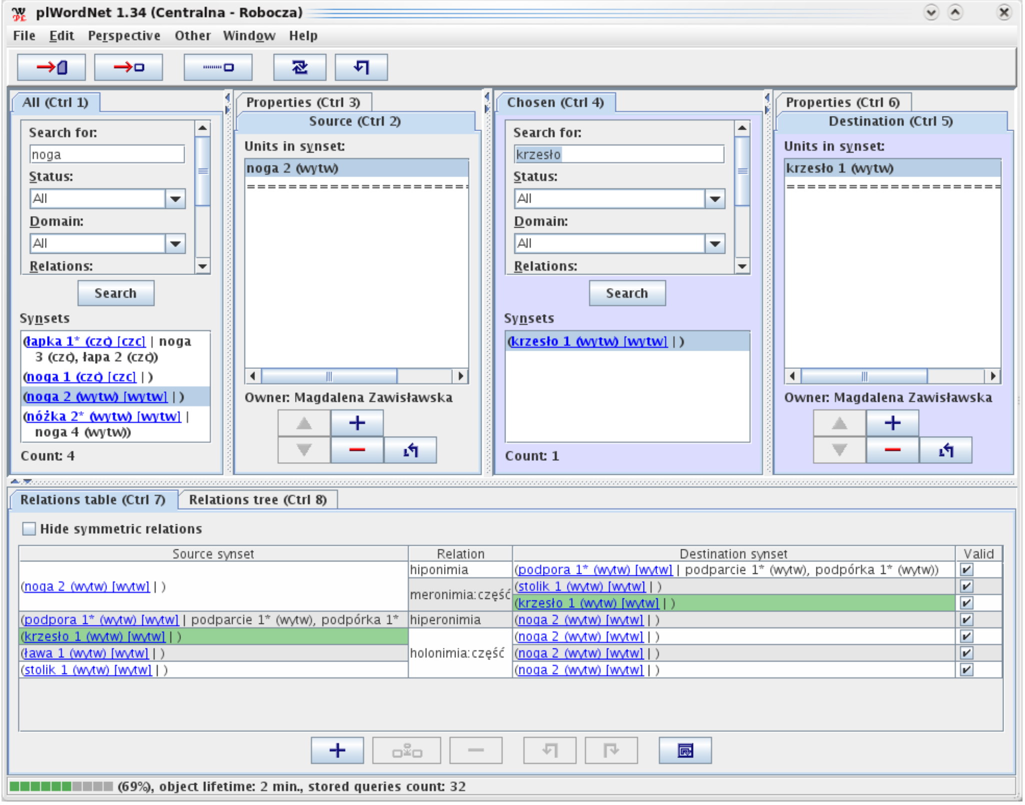
From the panel of the properties of the selected synset (the bottom right panel), the user can switch to the synset perspective, which is set at that moment to this synset as the *source synset*.

The five panels of the synset perspective (Figure 2.2) can be divided into the following groups:

- selection and editing of a *source synset* (two panels on the left side of the upper part) – the synset for which we are going to define a relation or whose relations we are going to browse and edit,
- selection and editing of a *target synset* (two upper right panels) of a relation to be defined,
- browsing of the existing relations (the bottom panel).

Two views of synset relations are possible: a tabular view (Figure 2.2) and a tree view (the hidden tab pane). According to the linguists' demands, the initial browsing-

[6] This is the place for adding glosses in the future.



(łapka 1* (czc) [czc] | noga 3 (czc), łapa 2 (czc))
(noga 1 (czc) [czc] |)
(noga 2 (wytw) [wytw] |)
(nóżka 2* (wytw) [wytw] | noga 4 (wytw))
Count: 4

(krzesło 1 (wytw) [wytw] |)
Count: 1

□ Hide symmetric relations

Source synset	Relation	Destination synset	Valid
(noga 2 (wytw) [wytw] |)	hiponimia	(podpora 1* (wytw) [wytw] | podparcie 1* (wytw), podpórka 1* (wytw))	☑
	meronimia:część	(stolik 1 (wytw) [wytw] |)	☑
		(krzesło 1 (wytw) [wytw] |)	☑
(podpora 1* (wytw) [wytw] | podparcie 1* (wytw), podpórka 1*	hiperonimia	(noga 2 (wytw) [wytw] |)	☑
(krzesło 1 (wytw) [wytw] |)	holonimia:część	(noga 2 (wytw) [wytw] |)	☑
(ława 1 (wytw) [wytw] |)		(noga 2 (wytw) [wytw] |)	☑
(stolik 1 (wytw) [wytw] |)		(noga 2 (wytw) [wytw] |)	☑

(69%), object lifetime: 2 min., stored queries count: 32

Figure 2.2: The synset perspective

only facility was extended with the editing of synset relations directly in this view. The browsing panel also allows the navigation along the graph of relations. The possibility of editing synsets directly in this perspective was introduced in order to facilitate the correction of the initial synsets. For example, it is possible to extract LUs from the source synset and to create a new hypernym synset.

Whenever the user wants to introduce an instance of a synset relation, the appropriate *substitution test* (Section 2.2 and Appendix A) is presented. The presentation of the tests was intended as a means of helping the consistency of the linguists' decisions. The test is shown before the user has added any new instance of a lexico-semantic relation. The test templates are defined, and can be edited, by coordinators in a dedicated window[7]. A test instance is generated from a *template* by instantiating it with the word forms of the tested LUs; we applied data collected from word forms in IPIC

[7] The set of functions available to the user depends on her role. For example, only coordinators can freely edit definitions of lexico-semantic relations, and may add new lexico-semantic relations.

processed by the morphological analyser *Morfeusz* (Woliński, 2006)[8]. The inflectional properties of word forms are specified in templates by the IPIC tagset codes. In keeping with the definition of the lexico-semantic relations as linking synsets (Section 2.2), a relation must be valid for any pair of LUs from both the source and the target synset. The substitution test window lets the user choose all possible pairs of LUs (from both synsets) and generate instances of the test. The mechanism of tests makes plWNApp different from other wordnet editing tools such as DEBVisDic, but we have not yet evaluated the influence of the substitution tests on the quality of plWordNet.

The structuring of plWNApp into two main screen-perspectives (stemming from the initial separation of broad synset construction and identification of relation instances) appeared to be incompatible with user expectations. The linguists signalled several times that a serious weakness of the application was the inability to show on one screen all synsets and all relation instances to which an LU belongs, especially in the case of relations that link LUs directly in plWordNet[9]. This request, however, is difficult to meet without assuming the availability of very large high-resolution monitors. Some linguists also strongly preferred keyboard interaction, and that discouraged further development of graph-based interface, which would require using the mouse. On the other hand, we found a two-perspective GUI quite successful in a recent extension of plWNApp to the WordNet Weaver (Section 4.5.3).

In order to improve the facilities of browsing LUs and the associated relation instances, we introduced an additional perspective (screen) of synset editing. This perspective, used for browsing synsets, has a layout similar to the LU perspective: a large list of synsets on the left (rich filtering possibilities), and on the right the tabular view of the selected synset relations, plus all synset editing panels. The screen and the LU perspective are synchronised: the filter setting and the selected synset or LU are transferred back and forth when switching. This facilitates browsing the LU relations of the LUs which belong to the given synset.

After the construction of the initial broad synsets, we proceeded to the next step: the formation of a net of lexico-semantic relations, that is to say, the proper core plWordNet. Two main problems arose. Synsets were too wide; they included not only the expected near-synonyms, but also hypernyms, co-hyponyms and even meronyms. Secondly, many synsets overlapped; many of such synsets belonged to the same domain, but their construction was separated in time and started with different LUs. The linguists had to extract hypernyms from the existing synsets and to divide synsets into more precise, smaller sets, using detailed guidelines including a number of substitution tests

[8] The tests were shown alongside all content-adding actions, so their content was often obvious. The users postulated the replacement of obligatory tests with an on-demand presentation of the instantiated test.

[9] The logistic and administrative circumstances of the project made it very hard to correct that, once the implementation has been largely completed.

for all types of relations (Section 2.2). The linguists were asked to consult several available dictionaries of Polish – (Dubisz, 2004) was selected as the basic one – and to follow the instruction (Derwojedowa et al., 2007). We discussed the main points of the instruction in Section 2.2. We present the substitution tests in Appendix A. The main problem in this phase of plWordNet development was that it took the linguists a long time to search for synsets to be connected. The initial synsets had been already created, so the preference was to link to them when introducing a new lexico-semantic relation instance. The existing synsets were not precise enough, and mixing construction with correction caused much trouble.

The division of LUs into domains was inevitably arbitrary, and their number was made low on purpose. Because the linguists were instructed not to cross domain boundaries[10], it was necessary to duplicate some LUs or even synsets. For example, we created two LUs for the lemma *budynek*; *budynek 1* 'building' in the domain `msc` (*locations*), and *budynek 2* 'building' in the domain `wytw` (*products*). We identified even more senses for the noun *odgłos*: *odgłos 1* 'hum, murmur' is in the domain `st` (*states*), *odgłos 3* '(any form of) voice' and *odgłos 4* 'tone' are in the domain `por` (nouns that name phenomena related to communication), *odgłos 2* '(any form of) sound' and *odgłos 5* 'echo' are in the domain `zj` (*natural phenomena*). For *budynek 1* and *budynek 2*, the separation of the two sense is fully justified: *budynek 1* has such hyponyms as *teatr* 'theatre', *ratusz* 'town hall' or *sejm* '≈parliament building', all combining the name of the building with its function; *budynek 2* is the hypernym for *kamienica* 'tenement house', *gmach* 'edifice', *hangar* 'hangar' – for the names of man-made objects.

The definition of as many as five senses of *odgłos* is much harder to justify. Such a large number of senses also makes the linguist's work harder. When constructing new relation links, she must in each case decide anew which sense should be selected as the target for a given new relation link. We could see clearly how many similar or even duplicate synsets had been constructed when, in the following step of semi-automatic expansion of plWordNet (Section 4.5.4), we dropped the restriction of one domain at a time – we began to allow relation instances to cross the domain boundaries at every level of plWordNet editing, The assignment of the new LUs introduced in the semi-automatic way (Section 4.5.3) was harder due to this phenomenon: there were several possible attachment points and the selection had to be based on subtle semantic distinctions. Nevertheless, domains of some granularity appear to be the most natural criterion for the distribution of the work among members of the linguistic team. Less constrained work of the linguists might be possible given much more flexible rules of

[10] In order to prevent two people from working on the same LU at the same time, coordinators were expected to take care of inter-domain relations and of merging parts of the structure created for different domains.

cooperation, supported by a more sophisticated mechanisms in the software tool (such as locking synsets and relation links while edited, version control and so on).

Several functions available in the synset perspective had been initially considered useful in dividing broad synsets into smaller synsets when constructing the network of lexico-semantic relations. For example, one could move selected LUs from one synset to another and join the two synsets by a relation, or move LUs between synsets. These options have been barely used at all. The linguists preferred to work on individual LUs. For example, during the editing of a synset all LUs were often deleted one by one, and new synsets including those LUs were created from scratch.

When we were designing plWNApp, we paid a lot of attention to such management issues as quality control and calculation of the amount of work every linguist performed. We assumed this typical work model: a draft version prepared by a linguist – evaluation done by a coordinator – error correction and final version prepared by the linguist. The practice showed that the requested calculations were very simple, given that no more than six linguists worked simultaneously. The mechanism of reporting errors was unused. The coordinators preferred making corrections on their own immediately after finding an error, because locating errors was also quite laborious.

2.5 The Final State of plWordNet Core

Manual construction of the core plWordNet took a substantial portion of the project's time (October 2005 – June 2008), but the three main activities – work on the linguistic foundations of plWordNet, the development of the wordnet editor plWNApp, and the editing of the wordnet database – went on in parallel, at least in the first year. The subsequent, much shorter, phase of semi-automatic expansion – based on the technology which we present in the next two chapters – resulted in almost doubling the core plWordNet to the present version plWordNet 1.0.

	Nouns	Verbs	Adjectives	*All*
Lemmas				
All	6085	3237	2617	11907
Monosemous	4531	2591	1913	8986
Polysemous	1554	646	704	2921
LUs	8544	4128	3781	16453
Synsets	5280	1595	2091	8966

Table 2.1: The size of the core plWordNet

Section 5.2 presents the state of plWordNet 1.0 in detail. As a reference point for the facts about plWordNet 1.0, we now show the core plWordNet in numbers.

A common practice in describing the size of a wordnet is to report the number of synsets it contains. This practice can be traced back to the original concept of PWN according to which a wordnet is an inventory of senses expressed by synsets. In our approach, LUs are the centrepiece of the wordnet. For many wordnet applications the numbers of lemmas (in the sense[11] introduced in Section 1.2) and the corresponding LUs (Section 2.1) are the most important characteristics of a lexical resource, representing its coverage and applicability. That is why we prefer to describe the size of the core plWordNet in lemmas and LUs, but for the sake of comparison we include the number of synsets, by part of speech. See Table 2.1.

	Including Monosemous Lemmas	Excluding Monosemous Lemmas
Nouns	1.405	2.586
Verbs	1.281	2.406
Adjectives	1.472	2.754

Table 2.2: Average polysemy in the core plWordNet

Around half of the lemmas employed in the construction of the core plWordNet were selected from the list of 10000 most frequent lemmas in IPIC. The remainder is due to the linguists' additions required to complete certain synsets, and to the attempts to translate the upper levels of PWN's hypernymy structure for the three parts of speech.

	Percentage of lemmas belonging to the n synsets [%]									
	1	2	3	4	5	6	7	8	9	≥ 10
Nouns	74.46	16.15	5.92	2.17	0.74	0.36	0.15	0.05	0.00	0.00
Verbs	80.04	14.21	4.17	0.99	0.40	0.19	0.00	0.00	0.00	0.00
Adjectives	73.10	15.74	6.61	2.71	0.99	0.08	0.27	0.15	0.15	0.20

Table 2.3: The number of synsets to which a lemma belongs in the core plWordNet

	Percentage of synsets including the n lexical units [%]									
	1	2	3	4	5	6	7	8	9	≥ 10
Nouns	65.91	19.50	7.91	3.82	1.38	0.62	0.36	0.13	0.17	0.20
Verbs	21.45	39.09	22.11	9.55	4.41	1.93	0.48	0.36	0.18	0.44
Adjectives	56.67	23.65	11.71	3.87	2.29	0.61	0.61	0.23	0.14	0.22

Table 2.4: Sizes of synsets in the core plWordNet

The division of lemmas into monosemous and polysemous, further illustrated in Table 2.2, was inspired by a similar practice of PWN (Miller et al., 2007). The polysemy statistics illustrate the general character of lemmas included in the core.

[11] Technically, a lemma is the basic morphological form of a given word form, produced by morphosyntactic disambiguation in context.

Relation	No. instances			
	Nouns	Verbs	Adjectives	*All*
Hypernymy	3966	315	134	4415
Holonymy	938	0	0	938
Meronymy	868	1	0	869
Troponymy	0	0	0	0
Antonymy	1035	147	1629	2811
Conversion	27	62	2	91
Relatedness	888	2378	1233	4499
Pertainymy	1002	190	304	1496
Fuzzynimy	389	41	425	855

Table 2.5: Instances of lexico-semantic relations in the core plWordNet

Table 2.3 completes the account of the polysemy of plWordNet. We show the distribution of different numbers of senses among lemmas of different parts of speech.

The core plWordNet includes mainly general LUs with vaguely specified meaning. That is why the ratio of large synsets is relatively high – see Table 2.4 – especially taking into account the assumed definition of synset.

Table 2.5 rounds out the detailed picture of the core plWordNet. It shows the statistics of instances of lexico-semantic relations. The numbers of holonymy and meronymy pairs differ because we set the value of reversibility to "–" (see the definition on page 26). So, inverse pairs are not created automatically: the linguist must make an explicit decision.

The tables we show in this section form a snapshot of plWordNet at a very special time of switching from the purely manual work to the semi-automatic expansion of plWordNet (based on the WordNet Weaver system, Section 4.5). Because all changes introduced in the second phase were verified by the linguists, there is no substantial difference in quality and trustworthiness between the plWordNet elements created before and after that time. Nevertheless, a comparison of the data presented in this section and the data shown in Section 5.2 for the final expanded version of plWordNet, offer an interesting comparison of the two different wordnet construction methods.

Chapter 3

Discovering Semantic Relatedness

3.1 Expectations

Manual construction of a wordnet on a large scale would normally require two resources that are always in short supply: time and money. The high cost is mostly due to the intensity of the linguists' labour. Thousands of decisions for thousands of LUs result in the magnitude of scale. Any reasonably complete wordnet is also expected to give a fairly exhaustive account of the LUs it represents and instances of lexico-semantic relations among those LU, though omissions are inevitable even if work is done vary carefully. In the core plWordNet, for example, only three most frequent senses of the lemma *zamek* were present by June 2008: 'castle', 'zipper' and 'door lock'. Omitted were the less frequent senses such as 'gun lock' and 'continuous puck possession during powerplay in ice hockey'. Extraction of lexico-semantic relations from large multi-domain corpora can help avoid missing senses. This is particularly important when expanding a wordnet to new domains and domain-specific vocabulary and senses.

Some researchers have argued that such technical support is indispensable:

> The sheer amount of knowledge necessary to shed light on the way word meanings mutually relate in context or distribute in lexico-semantic classes appears to exceed the limits of human conscious awareness and descriptive capability. (Lenci et al., 2001)

In some areas of Natural Language Processing, full task automation may be the norm. For wordnets, however, only semi-automatic construction is feasible. We argued that in Section 1.2 and Chapter 2.4. To sum up: first, a wordnet is treated as a dependable language resource, so it needs human control over its content. Second, contrary to our initial doubts, the meaning of the general LUs can be described properly on the basis of corpora – see the results of the extraction of semantic relatedness presented in Section 3.4 – but the contemporary automatic methods of synset extraction produce results far below human expectations (Section 3.5).

We will show how we propose to go about automating part of the development effort, but the ultimate responsibility for the shape of any wordnet, and especially the quality of its entries, rests with its authors. As a matter of general principle, a wordnet must be trustworthy. NLP researchers expect such lexical resources to be highly accurate; there may be even those who assume perfection and treat PWN as an authority. Without venturing into a discussion of the merits of such a stance, we adopt the tenet of trustworthiness. Thus, automated methods of extracting elements of a wordnet should at least deliver some form of intelligent browsing. This should draw the linguist's attention to selected LUs that could be linked to a given LU by a *wordnet relation*. Henceforth, we will mean wordnet relations as all lexico-semantic relations defined in wordnets, including PWN, EWN and plWordNet, with the emphasis on the latter.

Software tools in support of wordnet construction can be divided into four main classes, those which

- offer better *corpus-browsing* capability,
- *criticize* existing wordnet content,
- suggest possible *expansion* to existing wordnet content,
- suggest possible *substructures* of relations over LUs.

Corpus-browsing tools rely on the statistical analysis of a large corpus in search for distributional associations of LUs — this is discussed in Section 3.4 — or pattern-based extraction of relation instances — Section 4.

Tools that criticize an existing wordnet, termed *wordnet critics* here, can produce lists of relation instances missing in the wordnet, or mark already included relation instances as dubious. Wordnet critics can be based on the full range of methods: pattern-based, e.g. (Hearst, 1998), measures of semantic relatedness and hybrid combinations of the two; this is discussed in Section 4.5.

An automatic tool might suggest simple expansion of a wordnet, such as adding relation instances whose one element is already present in the wordnet. A more advanced suggestion might be the merging of a part of the wordnet with a subgraph of LUs linked by wordnet relations – in effect, a "sub-wordnet" – which has been extracted automatically and is connected to this part by at least a few relation instances. The simpler forms of expansion can be based on any combination of methods. To extract synsets automatically and link them to a wordnet by some relation is inconceivable without clustering LUs; we discussed this in Section 3.5.

3.2 Basic Division: Patterns versus Statistical Mass

Methods of the acquisition of lexico-semantic relations belong to two main categories that represent two paradigms of different origins; see (Matsumoto, 2003, Pantel and Pennacchiotti, 2006):

- pattern-based approaches,
- clustering-based approaches (also called *similarity-based approach* or *distributional similarity*, e.g. by Matsumoto (2003)).

Other authors note this division too, often implicitly, e.g. Widdows (2004).

Pattern-based approaches, e.g., (Hearst, 1992, 1998, Berland and Charniak, 1999), are based on applying manually constructed *lexico-syntactic pattern* to the identification of instances of lexico-semantic relations — LU pairs — in corpora. For example, that *linguist* and *scientist* can occur in the expression:

A linguist is a scientist who investigates human language…(Fromkin et al., 2000, pp. 3)

suggests that *linguist* is a hyponym of *scientist.* A lexico-syntactic pattern describes a class of language expressions by specifying partially their structures: selected lexical elements, types of constituents and syntactic relations. It is assumed that certain language constructions unambiguously indicate that pairs of LUs occurring in them are instances of certain lexico-semantic relations. Pattern-based approaches have relatively high precision for English (the situation for Polish is discussed in Section 4) but low recall. Substantial workload needed in the manual construction of patterns is reduced in methods that introduce partially automated extraction of patterns, e.g. (Morin and Jacquemin, 1999, Jacquemin, 2001, Morin and Jacquemin, 2004). The idea of full automation of pattern extraction resulted in the development of methods that combine extraction of generic patterns with statistical evaluation of their accuracy, e.g. (Brin, 1999, Agichtein and Gravano, 2000, Agichtein et al., 2001, Ravichandran and Hovy, 2002, Pantel et al., 2004, Pantel and Pennacchiotti, 2006). Section 4 discusses the pattern-based methods in detail.

Clustering-based approaches originate from the *Distributional Hypothesis* of Firth (1957) and Harris (1968) and characterise lexico-semantic relation of two LUs by the similarity of their corpus distributions, i.e. types of *contexts* in which they are used. One of the main results is the extraction of a *Measure of Semantic Relatedness* [MSR]. An MSR characterises semantic association between two LUs by some numerical value. That is to say, an MSR is a function:

$$L \times L \to R \tag{3.1}$$

where L is a set of lexical units, and R is a set of real numbers.

An MSR should assign higher values to semantically related pairs of LUs, so those LUs are grouped somehow. This implicit grouping gives the paradigm its name. Clustering-based methods are differentiated by the granularity and representation of contexts, as well as by the way in which the resulting values are calculated, e.g. (Ruge, 1992, Landauer and Dumais, 1997, Lin, 1998, Schütze, 1998, Widdows, 2004, Weeds and Weir, 2005), see Section 3.4. Section 3.5 discusses methods of clustering LUs given an MSR, which aim at the identification of groups of near-synonymous LUs, e.g. (Lin and Pantel, 2002, Pantel, 2003).

The two paradigms explore language data in corpora from significantly different perspectives: particular instances of specific language constructions in pattern-based methods, statistical regularities in clustering-based method. That is why the idea of their hybrid combination, especially for expanding a wordnet, arises quite naturally, e.g. (Caraballo, 1999, 2001, Girju et al., 2006). Hybrid approaches are discussed in Section 4.5.

3.3 Evaluation

Evaluation of the results of any acquisition method is crucial for its development, because most such methods depend on the values of several parameters. The methods differ in many aspects, so no single model of evaluation can be applied to all of them.

Those methods which produce list of LU pairs as associated by some target relation (such as hypernymy) can be evaluated by a manual inspection of the results — LU pairs. Those are pattern-based methods, see Sec. 4, and methods based on classifiers, see Sec. 4.5.1. Manual evaluation, however, is hampered by heavy workload required and by disagreement among evaluators. The former can be reduced by evaluating a representative sample and next ascribing the result to the whole set within the limits of the specified confidence level, see e.g. Section 4. The latter is inevitable, but the level of disagreement for lexico-semantic relations is low in comparison to the error of the automatic methods. Manual evaluation is a challenging task in the case of MSRs, see below, and other clustering-based methods, see Section 3.5.

Extracted lists of LU pairs can be automatically evaluated in a straightforward manner, via a comparison with an existing wordnet (or another thesaurus). We are, however, mostly interested in the behaviour of the method for the part of an evolving wordnet which has not been constructed yet. The result for the data not seen during evaluation can be only estimated. For example, we could often observe that the accuracy of our pattern-based method, when automatically measured against plWordNet, was stable or even decreased for subsequent settings. On the other hand, the manual evaluation was increasingly accurate — see Section 4.3. This problem appeared to be much harder for classifier-based methods, in which the differences between the results of the manual and automatic evaluation are especially visible, see Section 4.5.1.

Evaluation of the quality or effectiveness of an MSR is not trivial. Manual evaluation is barely feasible on a small scale. Not only are MSRs required to work for any pair of LUs, but also people are notoriously bad at working with real numbers. A linear ordering of dozens of LUs is nearly impossible, and even comparing two terms requires a significantly complicated setup (Rubenstein and Goodenough, 1965). Given a small sample of the lists of the most semantically related LUs to the given one, e.g., Table 3.11 and 3.12, people can easily distinguish a bad MSR from a good one; we must distinguish good MSRs from those that are merely passable from the perspective of support for linguists working on wordnet development.

We note three forms of MSR evaluation (Budanitsky and Hirst, 2006, Zesch and Gurevych, 2006):

- *mathematical analysis* of formal properties (for example, the property of a metric distance (Lin, 1998)),
- *application-specific evaluation*,
- and *comparison with human judgement*.

Mathematical analysis gives few clues with respect to the results of future applications of an MSR. Evaluation via an application may make it difficult to separate the effect of an MSR and other elements of the application (Zesch and Gurevych, 2006). A direct comparison to a manually created resource seems the least trouble-free. The construction of such resources, however, is labour-intensive even if it only labels LU pairs as similar (maybe just related (Zesch and Gurevych, 2006)) or not similar; this does not allow a fair assessment of the ordering of LUs on a continuous scale, as an MSR does.

Indirect comparison with the existing resources (Grefenstette, 1993) is another possibility. For example, one could compare an MSR constructed automatically and another based on the semantic similarity across the hypernymy structure of PWN. This is how the main approaches work – see (Lin, 1998, Weeds and Weir, 2005, Geffet and Dagan, 2004). Two list of the k LUs most similar to the given one – for example, one constructed from an MSR and one from a wordnet – are transformed to rank numbers of the subsequent LUs on the lists, and compared by the cosine measure. The drawback of such an evaluation is that we know how close the two similarity functions are, but not how people perceive an MSR. The evaluation also strongly depends on the wordnet similarity function applied. There are a number of such functions – see (Budanitsky and Hirst, 2006) – but many of them perform indifferently for a small wordnet without full-fledged hypernymy structure (like the core plWordNet that we had at our disposal during most experiments) or require synset probabilities. Moreover, wordnet similarity functions based on the hypernymy structure do not always work for verbs and adjectives, whose hierarchies tend to be quite limited. The similarity measure proposed by Mihalcea and Moldovan (1999) also does not apply in our case because plWordNet, like many other new wordnets, does not yet include glosses.

Automatic differentiation between words synonymous and not synonymous with a given LU is a natural application for an MSR, especially in the context of generation of suggestions for a linguist. In Latent Semantic Analysis [LSA] (Landauer and Dumais, 1997) the MSR constructed using a statistical analysis of a corpus (cf Section 3.4.2) was used to make decisions in a synonymy test, a component of the *Test of English as a Foreign Language* [TOEFL]. This gave 64.4% of hits. Turney (2001) reported 73.75% hits, and Turney et al. (2003) 97.5% hits; the latter practically solved the TOEFL synonymy problem. TOEFL is focused on humans, a big advantage for applications in MSR evaluation. On the other hand, it is manually constructed, hence its main drawbacks: limited size and fixed orientation on synonymy.

Freitag et al. (2005) proposed a *WordNet-Based Synonymy Test* [WBST], which seems to offer an interesting response to the limitations of TOEFL. WBST has been based on the use of PWN to generate "a large set of questions identical in format to those in the TOEFL". WBST is discussed in details in Section 3.3.1, but its two properties are worth emphasising now. First, it is larger and broader than TOEFL because it is automatically generated from a very large manually constructed resource. Second, with a change in the way of selecting question-answer pairs, a WBST-like test can evolve from a synonymy test to a test oriented toward wordnet relations or in the sense of (Mohammad and Hirst, 2006).

The best reported result for English nouns is 75.8% (Freitag et al., 2005). A slightly modified WBST was used to evaluate an MSR for Polish nouns (Piasecki et al., 2007a) with the result of 86.09%.

The evaluation of an MSR via a synonymy test shows the ability of the MSR to distinguish synonyms from non-synonyms. Since the MSR is the centrepiece of the application, the achieved results can be directly attributed to it. There was, however, a problem: WBST appeared to be too easy, as we show in Section 3.3.1. It is oriented toward testing the main distinction — closely semantically related versus unrelated — because the incorrect answers are selected randomly and on average they are semantically unrelated to the question and the answer. The usefulness of WBST is therefore limited with respect to its use in the development of more sophisticated MSRs focused on semantic similarity and wordnet relations.

In view of these findings, we have explored the possibility of generating more demanding automatic methods of MSR assessment, following the general idea of WBST. We proposed an *Enhanced WBST* [EWBST] which is precisely a template of WBST-like evaluation methods parameterised by the way in which *detractors*, i.e. false answers, are selected. We wanted its results to be easily interpreted by people and its feasibility tested on people. We also expected that it would pick the MSR that is a better tool for the recognition of lexico-semantic relations between LUs.

3.3.1 Wordnet-based synonymy test for Polish

The application of LSA to TOEFL data became unattractive as a method of comparing MSRs once the result of 97.5% hits has been achieved (Turney et al., 2003). Freitag et al. (2005) proposed a new test, WBST. It was seen as more difficult because it contained many more questions. An instance of the test consists of many — hundreds or even thousands — *question-answer pairs* [QA pairs]: $\langle q, A\rangle$, where $A = a_1, a_2, a_3, a_4$ and q, a_i are LUs included in the wordnet that underlies the test ((Freitag et al., 2005) used PWN 2.0). In each QA pair there is a_i, henceforth called the *correct answer*, such that there is a synset S in the wordnet and q, a_i belong to S. None of the other three a_j belongs to the same synset as q or as a_i. We will call such a_j detractors for the given QA pair. During evaluation, MSR generates values for the pairs $\langle q, a_i\rangle$, $a_i \in A$, expected to favour the correct answer against the detractors.

The WBST has been, amongst other applications, used to evaluate MSRs for Polish LUs (nominal, verbal and adjectival). The underlying resource was plWordNet, used in different development versions for different tests. Further in this section we discuss how the wordnet used influences the difficulty of the test.

The test had to be slightly modified. In plWordNet, many synsets have only 1–2 LUs, in accordance with the definition of the synset and usage of LUs as basic plWordNet entries, see Section 2.1. In order to get a better coverage of LUs by WBST questions, and not to leave LUs in singleton synsets untested, the direct hypernyms of LUs from singleton synsets were taken to form QA pairs[1] (Piasecki et al., 2007a). We named this modification the *WBST with Hypernyms* [WBST+H]. The inclusion of hypernyms in QA pairs did not make the test easier, as was shown in (Piasecki et al., 2007a).

plWordNet has been evolving from the early versions including fewer LUs, broader synsets with more vague understanding of near-synonymy (larger percentage of synsets with more than two LUs) and shallower hypernymy structure, to the present version of plWordNet expanded semi-automatically (Section 4.5.3), in which most synsets are narrow (1–2 LUs on average) and the hypernymy structure is significantly deeper. Having broader synsets puts in the same broad synset the LUs hard to distinguish using an MSR (they are very close in meaning). There is, therefore, no need to distinguish between their meaning during the test. In a version of plWordNet with narrower synsets, the same LUs may have already been separated into two different synsets usually not related by direct hypernymy. One can expect that narrower synsets obtained by dividing a broader one would be co-hyponyms. The hypernymy hierarchy deepened with the subsequent versions of plWordNet. This tendency was due to the

[1] In the case of adjectival LUs this technique has a limited application, because the number of hypernymy instances is very small in the case of adjectival synsets – only 142 instances (plWordNet from October 2008).

partitioning of broad synsets into (usually) hyponyms of the original one, and to the introduction of new lemmas in synsets which are hyponyms of the existing ones. This is illustrated in Table 3.1. We can observe the continuous decrease of the average synset size and the increase of the number of single-LU synsets.

plWordNet	12.2006	9.2007	6.2008	11.2008	plWordNet 1.0
Lexical units	11690	13164	16549	19620	26984
Synsets	5314	8045	9085	11880	17695
Singelton synsets	874	3745	5055	7660	12609
LUs in synset (average)	2.20	1.64	1.82	1.65	1.52

Table 3.1: Changes in synset structure during the development of plWordNet

In order to visualise the evolution of the WBST+H instances we used the best MSRs extracted for: nominal, verbal and adjectival LUs on the basis of the $\mathrm{MSR}_{GRWF(Lin)}$ algorithm, which will be discussed in Section 3.4. The same three MSRs were tested with different versions of WBST+H produced from different archival versions of plWordNet. The results appear in Table 3.2.

		WBST+H			EWBST		
plWordNet	PoS	Acc. [%]	Lemmas	QA	Acc. [%]	Lemmas	QA
12.2006	N	86.90	3661	10402	64.81	1780	4029
12.2006	V	81.34	2567	3905	—	—	—
12.2006	A	82.63	1547	3484	—	—	—
9.2007	N	85.99	3921	7522	66.15	3492	6512
9.2007	V	79.16	2567	4179	—	—	—
9.2007	A	84.48	1580	3530	—	—	—
6.2008	N	86.30	3816	6729	68.10	3391	5746
6.2008	V	75.29	2688	4734	—	—	—
6.2008	A	83.61	1567	2690	—	—	—
11.2008	N	88.14	5413	9486	69.75	5061	8689
11.2008	V	71.85	2677	5484	—	—	—
11.2008	A	83.26	1574	2814	—	—	—
plWN 1.0	N	87.60	9250	16826	73.28	8828	15832
plWN 1.0	V	71.06	2910	6340	—	—	—
plWN 1.0	A	81.53	1595	2875	—	—	—

Table 3.2: The accuracy of the MSRs based on the *Rank Weight Function* algorithm (Section 3.4) in relation to different tests and plWordNet versions (Lemmas – the number of lemmas in QA pairs, QA – the number of QA pairs for the given test); *plWN 1.0* refers to plWordNet, version 1.0

Examples of QA pairs taken from different WBST+H versions are presented in Figures 3.1 and 3.2.

Nouns		
Q:	diabeł (*devil*)	
A:	biolog (*biologist*),	gęganie (*(goose) cackling*),
	szatan (*Satan*),	wydech (*exhalation*)
Q:	pojazd kosmiczny (*spaceship*)	
A:	gałąź (*branch*),	**prom kosmiczny** (*space shuttle*),
	regulamin (*statute*),	znak rozpoznawczy (*distinguishing mark*)
Verbs		
Q:	królować (*reign (as a king)*)	
A:	nadążyć (*keep up*),	oderwać (*tear off*),
	panować (*rule*),	zauważyć (*notice*)
Q:	pragnąć (*desire*)	
A:	kompletować (*complete*),	**łaknąć** (*crave*),
	przystać (*agree, fit*),	uprzywilejować (*privilege*)
Adjectives		
Q:	dorosły (*adult*)	
A:	oryginalny (*original*),	**pełnoletni** (*of age*),
	przestarzały (*obsolete*),	złowieszczy (*ominous*)
Q:	nieprzenośny (*immobile*)	
A:	bryłowaty (*bulky*),	**stacjonarny** (*stationary*),
	weekendowy (*weekend*$_{adj}$),	żółtawy (*yellowish*)

Figure 3.1: Examples of WBST+H questions for plWordNet 1.0

plWordNet 12.2006	
Q:	nadzieja (*hope*)
A_1:	optymizm (*optimism*)
A_2:	przeświadczenie (*conviction*)
A_3:	otucha (*good cheer*)
A_4:	ufność (*confidence*)
A_5:	wiara (*faith*)
A_6:	szansa (*chance*)
A_7:	przypuszczenie (*supposition*)
plWordNet 1.0	
Q:	nadzieja (*hope*)
A_1:	otucha (*good cheer*)
A_2:	pokrzepienie (*fortification*)
A_3:	pocieszenie (*consolation*)

Figure 3.2: Examples of nominal QA pairs generated from different versions of plWordNet

Several sets of tests compared the results of MSR with human performance. Each time a subset of QA pairs was randomly selected from a complete WBST+H test and a group of native speakers of Polish were asked to solve the test. They were instructed to select for each question word only one answer, the closest in meaning to the question. There was no time limit in the task. Most participants were Computer Science students, but the LUs selected were mostly frequent units without technical senses, so the raters' background need not have influenced the results.

The first two tests for nominal LUs were generated from early versions of the core plWordNet:

- plWordNet from June 2006, 24 native speakers of Polish tested on 2 random subsets of WBST+H; a set included 79 QA pairs; the average score was 89.29%, and interjudge agreement within one set, measured by Cohen's kappa (Cohen, 1960), ranged between 0.19 and 0.47 (Piasecki et al., 2007a);
- plWordNet from March 2007, several native speakers of Polish, a random subset of WBST+H; the average result close to 100%.

The results of the second test showed the limits of WBST+H. Lacking a fuller version of plWordNet, we decided to define a more difficult test, WBST-style test to facilitate further work on MSRs for Polish nouns. This *Enhanced WBST* is presented in detail in the next section.

We also ran tests for verbal and adjectival LUs, both generated from the March 2007 version of plWordNet. Twenty raters solved each test of a hundred QA pairs. The participants' average scores appear in Table 3.3. The inter-judge agreement was measured by Fleiss's kappa, which accounts for agreement among many participants (Fleiss, 1971). The high value of kappa, supported by the manual evaluation of the test results, shows that the agreement was high, and the raters made similar errors. Examples of QA pairs appear in Figure 3.2. A comparison of the results of human raters on the verbal and adjectival QA pairs – 88.21% and 88.9%, respectively, with almost 100% for the nominal pairs – shows that the verbal and adjectival parts of WBST+H are more difficult for humans[2] and that one should expect lower results from the automatically extracted MSRs (Section 3.4.5).

In 2008, another WBST+H was generated for nouns, verbs and adjectives from the final version of the core plWordNet (June 2008). 80 LUs were selected randomly in 4 groups of 20 LUs for each range of LU frequency in the IPI PAN corpus (Przepiórkowski, 2004). We asked invited native speakers of Polish, mainly students of Computer Science, to solve the tests via dedicated Web pages. The results and the number of raters appear in Table 3.4.

[2] All three tests were generated from the same version of plWordNet.

PoS	Min [%]	Avg [%]	Max [%]	Kappa
Verb	84	88.21	95	0.84
Adjective	82	88.9	95	0.85

Table 3.3: Results of a manual WBST for Polish verbs and adjectives – the evaluation performed for (Broda et al., 2008) (May 2007)

	R	Min [%]	Max [%]	Avg [%]
Noun	29	73.84	96.24	86.64
Verb	50	57.54	90.04	81.84
Adjective	43	76.24	96.24	89.94

Table 3.4: Results of human raters in WBST+H tests generated from the final version of the core plWordNet (R — a number of raters for the given test)

It is misleading to compare the results in Table 3.4 with the almost 100% in WBST+H generated from the May 2007 plWordNet. The increase from 89.29% for June 2006 plWordNet to nearly 100% for May 2007 plWordNet was caused by the removal of many obvious errors in broad synsets of the early version of plWordNet. In many QA pairs of the former test, raters were misled by strange QA pairs occurring in the test. So, we can assume the level of almost 100% as the starting point. Considering this, when people solve the tests, we can observe a relation between the wordnet used and the difficulty of the WBST+H test opposite to what happens when MSR is applied: the results are slightly higher for new versions of WBST+H, see Table 3.2. The test results (produced for the same MSR) stayed approximately at the same level for the subsequent versions of the core plWordNet, and increased with the present version of plWordNet expanded semi-automatically with several thousand LUs (Section 4.5.4).

3.3.2 Enhanced WBST

In the WBST defined by Freitag et al. (2005) the elements of the answer set A not synonymous with Q are chosen at random from the whole wordnet. Thus, the difference in meaning between Q and the detractors is usually obvious to test-takers[3]. It also tends to be relatively easy for a good MSR, e.g. (Piasecki et al., 2007b). Our overall goal, however, was to construct an MSR that expresses clear preference for the wordnet relations (focused on semantic similarity in the sense of Mohammad and Hirst (2006) — Section 3.4.2). Such MSR could be used to automatically extract synsets, i.e. to

[3] The latest versions of the expanded plWordNet introduced more fine-grained distinctions between lemma senses. This made WBST+H more difficult for humans, as shown in Table 3.4 in relation to the previous test results discussed.

differentiate the LUs in a synset from all other LUs similar but not synonymous, among them co-hyponyms. Any such MSR must therefore distinguish closely related LUs, not only those with very different meaning.

In modifying the WBST+H test we assumed that we needed to construct the answer set A so that non-synonyms are closer in meaning to the correct answer a_i than it is the case in WBST+H. Obviously, they cannot be synonyms of either a_i or Q, but they ought to be related to both. We need to select the non-synonyms among LUs similar to s and to Q. In order to achieve this, we have decided to leverage the structure of the wordnet in the determination of similarity and to construct a *semantic similarity function* SSF_{WN} based on the plWordNet hypernymy structure:

$$SSF_{WN} : \mathbf{S} \times L \to R \tag{3.2}$$

where **S** is a set of synsets, L — lexical units, R — real numbers.

SSF_{WN} takes a synset S (e.g. including Q and a_i) and a lexical unit x (e.g. a detractor), and returns the semantic similarity value.

During the generation of the modified Enhanced WBST [EWBST], non-synonyms are still selected at random but only from the set of LUs broadly similar to Q and a_i. The acceptable values of $SSF_{WN}(S_Q, x)$ are lower than some threshold sim_t if the synset S_Q contains Q and a_i, and x is a detractor. We tested several wordnet-based similarity functions (Agirre and Edmonds, 2006), here implemented using plWordNet's hypernymy structure, and achieved the best result in a generated test with the following function:

$$SSF_{WN} = \frac{p_{\min}}{2d} \tag{3.3}$$

$p_{\min}$ is the length of a minimal path between two LUs in plWordNet, and d is a maximal depth of the hypernymy hierarchy in the current version of plWordNet. The similarity threshold $sim_t = 2$ for this function has been established experimentally. To achieve consistency between tests generated from different versions of plWordNet, we decided to set the sim_t to value corresponding to four arcs in hypernymy hierarchy.

The hypernymy structure of nouns in plWordNet does not have a single root, because in plWordNet we have not introduced any artificial common root nodes for all nominal LUs[4] Many methods of similarity computation require a root, however, so we have introduced a virtual one for the sake of the similarity computation, and linked to it all trees in the hypernymy forest.

We noticed that the random selection of LU detractors based any similarity measure tends to favour LUs in the hypernymy subtrees other than Q, if Q is located near the root. The number of LUs linked by a short path across the root is much higher than

[4] The same is the case for verbal and adjectival LUs, whose hypernymy structures are also partial and quite shallow.

the number of LUs from the subtree of Q which are located at a close distance to Q. The problem is especially visible for question LUs in small hypernymy subtrees with a limited number of hyponyms. The problem appears in the case of any similarity measure based on the path length, so we have heuristically modified the measure by adding a constant $\delta_R = 3$ to any path going across the virtual root. Lower values of δ_R gave no visible changes, while the higher numbers caused a large reduction of the number of QA pairs.

The difference in the level of difficulty between WBST+H and EWBST is illustrated in Figure 3.3 by an example problem generated by this method for the same QA pair: $\langle$majątek (*property, estate*), mienie (*property*) $\rangle$.

	EWBST	
Q:	majątek (*property, estete*)	
A:	lokata (*deposit, investment*),	**mienie** (*property*)
	obligacja (*bond, stock*),	wkład (*deposit, outlay*)
	WBST+H	
Q:	majątek (*property, estete*)	
A:	dzieciuch (*child, brat*),	**mienie** (*property*)
	rynsztok (*gutter*),	stryj (*uncle, father's brother*).

Figure 3.3: Example of the difference between EWBST and WBST QA pairs

Similarly to the tests performed for WBST+H, we have assessed the influence of the evolution of plWordNet on the MSR performance in EWBST. The same algorithm of extraction was used as in the case of the former experiments: $\mathrm{MSR}_{GRWF(Lin)}$ discussed in Section 3.4. Only a MSR for nominal LUs was built, because EWBST depend strongly on the hypernymy structure. The same MSR was tested with different versions of EWBST produced from different archival versions of plWordNet. The results are presented in the joint Table 3.2. Examples of EWBST test instances are presented in the Fig. 3.4.

We can observe for EWBST results a similar tendency as for WBST+H. The EWBST test becomes slightly easier as plWordNet evolves (we hope that it improves): from 64.81% to 69.75%. For EWBST, however, the increase is continuous with each version of plWordNet – WBST+H shows a larger difference only between the final core plWordNet and the expanded version. The increase for EWBST may be due to the deepening of the hypernymy structure. There are two possible reasons for the observed changes of the MSR results in relation to different tests. The introduction of many specific LUs in the expanded version of plWordNet made both tests easier: specific LUs are easier to distinguish. EWBST was getting easier with the deepening hypernymic structure, as LUs grouped earlier in large vague synsets were distributed

	EWBST, Nouns, plWordNet 12.2006	
Q:	aromat (*aroma*)	
A:	**bukiet** (*bouquet*),	fetor (*stench*),
	smrodek (*stink (diminutive)*),	smród (*stink*)
	EWBST, Nouns, plWordNet 09.2007	
Q:	aromat (*aroma*)	
A:	**bukiet** (*bouquet*),	fetor (*stench*),
	powódź (*reason*),	upał (*heat*)
	EWBST, Nouns, plWordNet 1.0	
Q:	aromat (*aroma*)	
A:	**bukiet** (*bouquet*),	piorun (*thunderbolt*),
	widmo (*phantom*),	zadymka (*snowstorm*)
	WBST+H, Nouns, plWordNet 1.0	
Q:	aromat (*aroma*)	
A:	**bukiet** (*bouquet*),	faworyzowanie (*favouring*),
	harówka (*drudgery*),	matematyka (*mathematics*)

Figure 3.4: Examples of QA pairs with detractors generated from different versions of plWordNet for the same QA pair

along the structure and less frequently drawn as detractors. The QA pairs generated from broad synsets were often vaguely semantically related and were harder for both tests to differentiate from the question-detractor pairs, which were often also vaguely related.

We also tested raters' performance on EWBST for the needs of future comparisons with the performance of the automatically extracted MSRs. During the first experiment, an example EWBST test generated from the March 2007 plWordNet was given to 32 native speakers of Polish, all of them Computer Science students[5]. The test consisted of 99 QA pairs. All LUs in the test were selected from 5706 single-word noun LUs in plWordNet. In the set of question LUs, 42 LUs occurred more 1000 times in the IPI PAN corpus (Przepiórkowski, 2004). This subset was distinguished in the test, because such LUs are also the basis of the comparison with the results achieved in (Freitag et al., 2005).

For all QA pairs the result was 70%, with the 61.62% minimum, 78.79% maximum and $\sigma = 4.07\%$ standard deviation from the mean. For the subset consisting of frequent LUs, the average result was 63.24%, with the minimum 52.38%, maximum 73.81% and $\sigma = 5.37\%$.

[5] As in experiments with WBST+H, this bias in the background should not influence the results, because the test was composed from plWordNet which at present includes only general Polish vocabulary.

The results, as expected, are much lower than those achieved in WBST+H tests. We were surprised that the results the raters had for the frequent LUs were significantly lower than for all LUs. It is likely that more frequent lemmas are at same time more polysemous, and that makes them more difficult to distinguish from other similar lemmas. The results for frequent LUs are lower, but at the level similar to the results for all LUs.

In 2008, in parallel with the new WBST+H versions, we generated a version of EWBST (for nominal LUs) based on the May 2008 plWordNet version. It included 80 LUs selected randomly in 4 groups of 20 LUs for each range of LU frequency in IPIC. Again, native speakers of Polish, mainly students of Computer Science, solved the tests via a dedicated Web page. The results and the number of raters are presented in Table 3.5.

Raters	Min [%]	Max [%]	Avg [%]
30	52.54	81.24	71.34

Table 3.5: Results of human raters in EWBST (for nominal LUs) generated from the final version of the core plWordNet

3.4 Measures of Semantic Relatedness

3.4.1 The distributional hypothesis and its consequences

Harris (1968) in his statement of the Distributional Hypothesis expressed a strong belief that there is a direct relation between the observed use of language expressions and their meaning (cited after to (Sahlgren, 2001)):

> The meaning of entities, and the meaning of grammatical relations among them, is related to the restriction on combinations of these entities relative to other entities.

Entities – language expressions or lemmas[6] (Section 1.2) that occur in text – interact via complex syntactic and semantic relations. The occurrence in text of a particular language expression is limited by constraints. Constraints are imposed by the co-occurring language expressions and the instances of grammatical relations induced by this coincidence. Each occurrence of a language expression can therefore be described by a set of lexicalised constraints. Harris contends that two language expressions with similar sets of lexicalised constraints across their occurrences have a similar meaning.

[6] Lemmas are more convenient in the case of inflectional languages: they help reduce the complexity caused by a large number of word forms.

LUs manifest themselves in text by occurrences of language expressions. That is why we can also extend the application of this hypothesis to the meaning of LUs.

The Distributional Hypothesis allows one to assess the commonalities between LU meanings by measuring the similarity of contexts in which they occur (via language expressions). Grammatical relations are recognised in text mostly exact to some degree of accuracy, so in the general case we should rather talk about measuring the strength of semantic relatedness between LUs – not the exact semantic relations between them. The semantic relatedness that is a correlate of the likelihood that two LUs can occur in the same type of contexts.

A *Measure of Semantic Relatedness* [MSR], briefly discussed in Section 3.2, is a function that assigns a real value to the semantic relatedness of two LUs by comparing the descriptions of their distribution across different contexts in the corpus.

High recall is an intrinsic property of an MSR. An MSR finds a value of the strength of relatedness for almost any pair of LUs. Moreover, for a given LU x and a large enough value of k one can expect many LUs related to x by one of the PWN relations among k LUs most semantically related to x – henceforth, we will denote this set of LUs by MSRlist$_{(x,k)}$. In practice, however, we mostly see a low accuracy of MSRs measured as the cut-off precision of the MSRlist$_{(x,k)}$ list calculated in comparison to relation instances extracted from a wordnet (Section 3.3) for a fixed value of k, such as 20. See the result of the experiments later in this section. Nevertheless, despite the expected problems with accuracy, but due to the expected high recall, our goal for the first step of constructing tools for semi-automatic expansion of plWordNet was to build an MSR for Polish with a relatively high accuracy with respect to the core plWordNet. We planned to achieve this by working with a very large corpus – to increase the number of examples of LU use – and by using rich description of contexts based on the analysis of morphosyntactic dependencies among LU occurrences. We expected to extract an MSR more focused on semantic similarity, which returns a large percentage of LUs associated with x by synonymy or hypernymy among MSRlist$_{(x,k)}$ for some LU x and a small value of k. The idea is to let the linguist browse the whole MSRlist$_{(x,k)}$ comfortably. Preliminary experiments also suggested that linguists might not accept less than 50% of correct instances of lexico-semantic relations on the list of suggestions.

3.4.2 Context and its description

The construction of an MSR requires two decisions first: on the *context size* (or *granularity* of the LU meaning description) and on the types of *constraints* used as context description. The decisions are correlated. For example, with context that exceeds sentence boundaries, the description cannot be based only on lexico-syntactic relations (most syntactic relations do not hold outside a sentence). Two main lines of work emerge in the literature – MSR extraction based on:

- *text windows* – the context is a window (a whole document in some special cases); co-occurrence with particular LUs serves as constraints;

- *lexico-syntactic constraints* – the context is a sentence, clause or phrase; lexico-syntactic relations serve as constraints.

Mohammad and Hirst (2006) write that this distinguishes between measures of *semantic relatedness* and *semantic similarity*, but we feel that intermediate methods are quite conceivable. For example, one can combine lexico-syntactic constraints with co-occurrences in the description of context. So, there is a continuum of methods with these two extremes.

In the seminal paper on *Latent Semantic Analysis* [LSA] (Landauer and Dumais, 1997), a context is simply the whole document (longer documents were truncated to a predefined size). The created *co-incidence matrix* (also called *co-occurrence matrix*) describes nouns[7] by the frequencies of their occurrences across documents. Rows correspond to nouns (60 768), columns to documents (30 473), and a cell $\mathbf{M}[n_i, d_j]$ stores the number of occurrences of the noun n_i in the document d_j. The initial cell values are then *weighted* by the *logent* function (Section 3.4.4) and the whole matrix is transformed by *Singular Value Decomposition* (SVD) (Berry, 1992) to a matrix of reduced dimensions. The SVD transformation not only improves the efficiency by reducing the row size but also – much more important – emphasises relatedness[8] between particular nouns or its absence. The final MSR value is calculated by comparing, using the cosine measure, rows of the reduced matrix that describe particular LUs. The relatively good result of 64.4% achieved in the Test of English as a Foreign Language (discussed briefly in Section 3.3) may have been due to the high quality of the corpus: *Grolier Encyclopedia*.

In order to overcome the corpus size restriction induced in LSA by the application of SVD, Schütze (1998) proposed a method called *Word Space*. A text window moves across documents. At each position MP of the window, statistics are collected: co-occurrence of a word in the centre of the context with a number of *meaning bearers* (selected general words). Turney (2001) used the Altavista search engine to search for co-occurrences of LUs in millions of documents on the Internet and thus to calculate an MSR.

Experiments performed on Polish data (Piasecki and Broda, 2007) suggest that text-window contexts described by LU co-occurrences result in MSRs that produce

[7] Only noun word forms were described in the experiment of Landauer and Dumais (1997).

[8] Landauer and Dumais (1997) wrote about "similarity", but in keeping with the earlier remarks we prefer to talk about semantic relatedness, because LSA is a typical text-window MSR.

broader semantic associations among LUs[9]. Such contexts tend to extract semantic relatedness *sensu largo* rather than (more desirable) closer semantic similarity.

In approaches based on lexico-syntactic constraints, a target LU is described by instances of its lexico-syntactic relations with particular LUs. As an example, for the noun *bird* we find the constraint `subject_of(`*sing*`)` met in texts. Hindle (1990) used a deterministic parser and analysed relations of nouns with verbs as subjects and objects. Two measures, subject similarity and object similarity of two nouns in relation to a given verb, were calculated from the collected frequencies. The final MSR value for a pair of nouns was defined as a sum of both similarities across all verbs. In defining MSR for 26742 nouns, Hindle used only 4789 verbs for which at least one sentence or clause structure (274613 in total) was recognised by the parser. Lexico-syntactic constraints were applied for the construction of MSRs also by Ruge (1992), Grefenstette (1993), Widdows (2004), Weeds and Weir (2005).

Lin (1998) applied a shallow dependency parser, *MiniPar* (Lin, 1993), to the preprocessing and identification of syntactic dependencies that involve nouns. The number of different syntactic relation utilised for the MSR computation is not given; MiniPar recognises several hundred syntactic dependency relations, about 200 of which describe dependency links involving noun phrase heads. Examples in (Lin, 1998) suggest that many different relation were used in defining lexico-syntactic constraints. The correlation of the MSRlist$_{(x,k)}$ list generated from Lin's MSR with the MSRlist$_{(x,k)}$ list generated on the PWN-based similarity appeared to be much higher than the correlation with lists generated from the MSR proposed in (Hindle, 1990). The result showed that the use of a large set of syntactic dependencies, not only based on the subject and object relations, improves the MSR.

In the experiments on Polish data, we observed progress in WBST+H with the addition of constraints of different types. For example, here are the observations in the experiments performed for (Piasecki et al., 2007b): while the MSRs based on the individual constraints expressing only adjectival modification and noun coordination achieve 88.65% and 76.85%, respectively, an MSR based on the combination of both constraints achieves 90.92% in WBST+H. We also made a comparison of MSRs constructed as described in (Piasecki and Broda, 2007):

- LSA applied to a subcorpus of the IPI PAN Corpus [IPIC] (Przepiórkowski, 2004) including 185066 documents from a daily Polish newspaper – 58.07% in WBST+H generated from the core plWordNet,

[9] The same tendency could be observed in the similar experiments performed on a corpus of 584 million token (the joint corpus, Section 3.4.5 and plWordNet from November 2008. We compared two MSRs extracted for nominal LUs (13285, described in Section 3.4.5): one based on lexico-syntactic constraints and another on pure co-occurrence in the text window of ±5 tokens. The results achieved in WBST+H test, 88.14% and 75.20% respectively, and 67.95% and 58.86% in EWBST, seem to support the claim that the use of text-window contexts results in less precise discrimination of LU meanings.

- co-occurrence with adjectives in a very small text-window (±2) – 74.16% in the same WBST+H,
- and morpho-syntactic constraints describing modification by adjective (Section 3.4.3) – 81.15%.

The experiments showed that an MSR for Polish which takes into account syntactic relations, even limited to adjectival modification, more accurately differences between semantically similar and dissimilar nouns. So, an MSR based on lexico-syntactic constraints is more useful for wordnet expansion than an MSR based on text windows. There is, however, an obvious drawback in the premises of this argumentation: the corpus used in the first case was much smaller than the whole IPIC used in the other two experiments. Still, it is hard to predict what would happen, because the first experiment explored the limits of technical possibilities of SVD – we could not process any larger matrix. Motivated by the result of the third experiment, we decided to focus on the constraint-based MSRs.

Many methods have been proposed for MSR extraction, but they all contain four general steps, more or less clearly delineated.

1. *Corpus preprocessing* – typically up to the level of shallow syntactic analysis.
2. *Co-incidence matrix construction* – in which rows correspond to LUs being described and columns to features.
3. *Matrix transformation* – a possible reduction of size and/or combination of feature *weighting* and *selection*.
4. *Semantic relatedness calculation* – LU descriptions are compared by the application of the assumed measure of similarity between row vectors.

The following sections discuss the steps and the corresponding results for Polish.

3.4.3 Preprocessing based on morphosyntactic constraints

Landauer and Dumais (1997) used word forms collected from a corpus as elements described by MSR(LSA). Such a strategy is doomed to failure in the case of an inflectional language like Polish – there are too many word forms and word forms of the same lexeme have the same meaning but can have different distributions. A natural strategy would be to transform all word forms into their basic morphological word forms. There is, however, much ambiguity. For example, the word form *mam* can represent three different lexemes with three different base forms:

1. *mama* (mam = mom$_{case=gen,num=pl}$),
2. *mieć* (mam = have (possess)$_{person=1st,num=sg,tense=present}$),
3. *mamić* (mam = delude$_{imperative}$).

In order to disambiguate base form assignment, we applied the morphosyntactic tagger TaKIPI (Piasecki and Godlewski, 2006). The accuracy of the base form identification by TaKIPI is 99.31% (Piasecki and Radziszewski, 2009), as measured in relation to the manually disambiguated part of IPIC.

MSR extraction methods based on lexico-syntactic constraints assume that the corpus has been preprocessed by a parser. There is no available parser or shallow parser for Polish, which could be used for this task: *Swigra* (Woliński, 2005) is a deep parser that produce many possible detailed analyses for a sentence, the dependency parser of Obrębski (2002) also returns several analyses for a sentence, and the *Poleng* parser (Graliński, 2005) is a commercial product, whose version available for the plWordNet project caused problems with interpreting the output format[10].

Faced with the lack of a suitable parser, we considered the morphological information encoded by Polish word forms. It has turned out to be rich enough for use in a tool to replace a parser. *Lexico-morphosyntactic constraints* as context descriptors help identify semantically relevant association between a target LU and other LUs in the lexicon. In Polish, associations among language expressions very often depend on the morphosyntactic characteristics of their constituents, such as gender/number/case agreement between an adjective and a head noun. In an inflectional language like Polish, the morphosyntactic description of word forms (rather than word order) delivers most of the structural information. For example, an adjective and a noun which are constituents of the same noun phrase can occur in both possible orders[11] but the agreement is necessary. Morphosyntactic associations are also simpler to recognise, since this requires only a tagger and a constraint representation formalism. Morphosyntactic taggers have been created for most European languages; in our experience, a constraint language interpreter can be constructed for a given language in a few person-weeks.

The JOSKIPI language, originally introduced as the language of tagging rules in TaKIPI, was used to implement morphosyntactic constraints. Selected elements of JOSKIPI will be presented as we discuss the constraint examples later in this section. For a detailed description, see (Piasecki, 2006, Piasecki and Radziszewski, 2009). In general, the expressions are used to recognise potential associations between a target LU occurrence and occurrences of other LU in the given sentence. Each constraint is based on a template that has a marked place for a LU – a *lexical element*. A set of concrete constraints is generated from a list of lexical elements predefined for the

[10] It was designed as an internal module of a Machine Translation system.

[11] Except some fixed collocations.

given constraint template. Lists of lexical elements can be freely defined, but they are mostly acquired directly from corpora, e.g. a list of all adjectives occurring in a corpus. A co-incidence matrix based on constraints has the following scheme:

$$\mathbf{M}[w_i, c_t(x)] \tag{3.4}$$

w_i is one of the target LUs, $c_t(x)$ is the template c_t lexicalised with the LU x.

A cell $\mathbf{M}[w_i, c_t(x)]$ stores the number of occurrences of w_i in the corpus which met the lexico-morphosyntactic constraint $c_t(x)$. In order to simplify the description, we will refer to the constraints as *features* which describe the target LUs semantically, and to the cells as *feature values*.

A constraint $c_t(x)$ is activated for the given occurrence of w_i during matrix construction when x occurs in the given sentence. Constraints are applied to morphosyntactically tagged text. They can test token annotations in some positions referred to by offsets to the context match centre (the position of w_i) and can iterate across the whole sentence. All JOSKIPI-based constraints return Boolean values that depend on the given w_i position and the surrounding sentence.

Constraints of several types were tested for the description of nouns (Piasecki et al., 2007b). In the end, four types were selected as producing an MSR with the best results in WBST+H:

`AdjC` – modification by *a specific adjective* or *a specific adjectival participle*,

`NcC` – co-ordination with a *a specific noun*,

`NmgC` – modification by *a specific noun* in the genitive case,

`VsbC` – occurrence of *a specific verb* for which a given noun can be its subject,

The `AdjC` constraint presented in a schematic form in Figure 3.5 is a example of a constraint strongly based on morphosyntactic agreement – here on case, number and gender. Such constraints are relatively easy to recognise and have high accuracy in recognition, see Table 3.6, discussed later. In `AdjC`, first we are looking for a particular adjective or an adjectival participle (specified by the base form) to the left of the target LU N in the position `0`:

- the `llook` operator implements searching for tokens that meet the condition given as its last argument,
- `$A` is a variable used for iteration (all variable names start with '$'),
- `in` and `inter` are set operators of inclusion and intersection,

```
or(
  and(
    llook(-1,-5,$A,and( in(flex[$A],{adj,pact,ppas}),
                        inter(base[$A],{"particular base form"}),
                        agrpp(0,$A,{nmb,gnd,cas},3)
    )),
    or(
      only($A,-1,$Ad, in(flex[$Ad],{adjectival and adverbial grammatical classes,
                                     numerals and punctuation})),
      and(
        in(cas[0],nom,acc,dat,loc,inst,voc),
        there is no other verb then "być" between -1 and $A positions
        not(
          llook(-1,$A,$S,and(
                          in(flex[$S], {nominal grammatical classes}),
                          in(cas[$S],{nom,acc,dat,loc,inst,voc}),
                          not( llook($S,$A,$P,equal(flex[$P],{prep})) )
          ))
        )
      )
    )
 ),
 a symmetrical condition for the right context
)
```

Figure 3.5: Parts of a lexico-morphosyntactic constraint which describes nominal LUs via adjectival modification (AdjC)

- flex returns a *grammatical class*[12] of the specified token,
- adj, pact, ppas are mnemonics for grammatical classes of adjective and two adjectival participles,
- agrpp(0,$A,nmb,gnd,cas,3) is an operator that tests agreement between two specified positions and according to the given list of grammatical categories[13].

After the lexical element A has been found and its position stored in $A, we need to test if no tokens between N and A make the modification of N by A impossible. For example, A may belong to a different noun phrase than N, so agreement is accidental. In the following steps of the constraint AdjC, then, we test two situations that validate the modification:

[12] In the tagset of IPIC (Przepiórkowski, 2004), word forms are divided into 32 grammatical class, a division more fine-grained than parts of speech; this is motivated largely by morphological, derivational and syntactic properties of word forms.

[13] The last parameter has a technical meaning for more advance uses of agrpp. It describes the number of categories.

1. only adjectival words, adverbial words, numerals or punctuation occur between N and A (`only` iterates across tokens and applies the specified condition),

2. there is an occurrence of *być* (*to be*) between N and A (an attributive use of *być*) and there is no other noun between N and A, which could be the real head for the modification by A.

The second condition is also constrained by the requirement of N not being in the genitive case: a noun in genitive can also be a modifier, so predictions may be less accurate.

`VsbC` is based on the nominal case of the target LU N – a potential subject and the agreement on number and gender or only gender, depending on the verb form, between N and a lexical element V – a potential predicate for N. Such an agreement is too weak evidence, so the presence of any other potential subject N' intervening in this possible association is tested. N' can occur at any position in the sentence, so a range of possibilities is tested.

`NcC` depends only on the case of the two nominal LUs which may be coordinate: the target LU N and the lexical element M. We identify M's position and we check the equality of case values of N and M. Next, the tokens occurring between N and M are also tested for representing only a limited number of grammatical classes.

```
and(
  rlook(1,end,$B,and(
                   in(flex[$B],{nominal grammatical classes}),
                   equal(base[$B],{particular base form}),
                   equal(cas[$B],{gen})  )),
  only(1,$-1B,$Ad, or(
                    in(flex[$Ad],adverbial grammatical classes),
                    and(
                      in(flex[$Ad],{nominal grammatical classes}),
                      equal(cas[$Ad],{gen})
                    ),
                    and(
                      in(flex[$Ad],{adverbial grammatical classes and numerals}),
                      agrpp(0,$Ad,{nmb,gnd,cas},3)
                    )
)) )
```

Figure 3.6: Parts of a morphosyntactic constraint which describes nominal LUs via the modification by a nominal LU in the genitive case (`NmgC`)

`NmgC`, presented schematically in Figure 3.6, identifies modification by a specific noun in genitive, which represents an ambiguous or even vague semantic relation. The constraint does not depend on any morphosyntactic agreement; that makes it hard to recognise properly. We have, however, significantly limited the range of constructions where this constraint is met, so the achieved accuracy is relatively good – see Table 3.6. Moreover, in testing the presence of adjectival words between the target LU and the lexical element, we refer to agreement for more accurate recognition. Modification by a nominal lemma in genitive clearly refers to the lexical meaning of the modified nominal. `NmgC` had a positive influence on the accuracy in the experiments – see (Piasecki et al., 2007b, Piasecki and Radziszewski, 2009).

Verbal LUs are described in plWordNet not in terms of subcategorisation frames, but by the semantic and lexical relations. So instead of recognising syntactic frames[14], we applied morphosyntactic constraints in a way similar to the description of nominal LUs. The description of occurrences of verbal LUs comprises four templates of morphosyntactic constraints (the lexical elements have been italicised):

`NSb` – a particular *noun* as a potential subject of the given verb,

`NArg` – a *noun* in a particular case as a potential verb argument,

`VPart` – a present or past participle of the given verb as a modifier of some *nominal LU*[15],

`VAdv` – an *adverb* in close proximity to the given verb.

`NSb` is a symmetrical to the `VsbC` constraint applied to nominal LUs. Now nominal LUs are the lexical elements searched for. The `NArg` template is parametrised by two values: a case value (the nominative value is excluded as covered by `NSb`) and a nominal lexical element. Because there is no agreement between a verb and its argument and we had no description of verb subcategorisation frames for Polish, the `NArg` implementation is very straightforward. Having the verb in the centre (position `0`) we are looking for the first occurrence of the given lexical element in the given case unless it is separated by an occurrence of another verb (when we cannot disambiguate the attachment). `VPart` explores the common use of present and past participles as adjectival modifiers of nominal LUs. Verbs are described via their occurrences as participles and lexical elements are the modified nominal LUs. The constraint is very similar to the `AdjC` constraint for nominal LUs. For the `VAdv` constraint we test the presence of an lexical elements – an adverb – at the two closest positions to the left or right. Adverbs have no grammatical categories except degree, so only distance can be considered.

MSRs for adjectives were constructed as a by-product of larger projects in (Hatzivassiloglou and McKeown, 1993, Freitag et al., 2005). Extraction of distributional

[14] This might be very difficult due to the lack of a shallow parser.

[15] A subtle agreement test and additional structural conditions distinguish such pairs from verb-complement pairs.

features was also discussed in (Lapata, 2001, Boleda et al., 2004, 2005), but applied in the semantic classification of adjectives. We have identified three types of constraints as the potential semantic descriptors of adjectives:

`ANmod` – an occurrence of a particular *noun* as modified by the given adjective,

`AAdv` – an *adverb* in close proximity to the given adjective,

`AA` – the co-occurrence with an *adjective* that agrees on case, number and gender as a potential co-constituent of the same noun phrase.

`ANmod` is symmetrical to the `AdjC` constraint used for nominal LUs, but this time lexical elements are nouns instead of adjectives. `AAdv` is very similar to `VAdv`: lexical elements are adverbs and we test the presence of an adverb in a distance not greater than 2. The implementation of `AA`, where lexical elements are adjectival LUs, has been based on the scheme of `ANmod`, but we are looking for an occurrences of another adjectival LU which agrees on case, number and gender and which can be a co-modifier of the same nominal LU.

The latter feature was advocated by Hatzivassiloglou and McKeown (1993) as expressing negative semantic information: only unrelated adjectives can sit in the same noun phrase. Our corpus data (collected from IPIC), however, suggest that it is too strong a bias. In addition, our AA constraint also accepts coordination of adjectives, and then related adjectives can co-occur in a noun phrase. In the end, we used the AA feature in a positive way, just like the other features. Features of all three types, weighted and filtered by the RWF weight function discussed in Section 3.4.4, were used in the discovery of contexts of occurrences of particular adjectives.

The AA constraint was applied in two different ways:

- as part of a joint large matrix together with the two other constraints: different parts (columns) of row vectors generated by different constraints, but the matrix processed as a whole – this usage is encoded ANmod+AAdv+AA in Table 3.13,
- two separate matrices were created: one joint for ANmod+AAdv and another for AA only.

In the second situation, the semantic relatedness values were calculated separately on the basis of both matrices separately processed and next linearly combined (Broda et al., 2008):

$$\begin{aligned} &MSR_{Adj}(l_1, l_2) = \\ &\alpha\ MSR_{ANmod+AAdv}(l_1, l_2) + \beta\ MSR_{AA}(l_1, l_2) \end{aligned} \tag{3.5}$$

The values of the coefficients were selected experimentally; $\alpha = \beta = 0.5$ gave the best results.

During the experiments performed by (Broda et al., 2008), a linear combination of separate matrices, that is, a linear combination of two MSRs, gave better results than the joint matrix ANmod+AAdv+AA. However, as the issue of extracting MSRs on the basis of the combination of separate matrices still requires more in depth research, we do not present here a repeated experiment of this kind.

The results of the manual evaluation of the constraints for nominal LUs, presented in (Piasecki and Radziszewski, 2009), appear in Table 3.6. For each constraint template and the appropriate list of lexical elements, the total number of matches in IPIC was calculated and based on that a sample of matches was randomly drawn. Each match of the lexicalised morphosyntactic constraint in the sample was extracted as a triple: the sentence, the described LU and the lexical elements. The positions of both expressions in the sentence were marked. The task of the evaluator (one of the co-authors) was to analyse if the relation described by the constraint holds for the given pair in the given sentence. The sample sizes were chosen according to the method described in (Israel, 1992), in such a way that the results of the sample evaluation can be ascribed to the whole set with a 95% confidence level.

	Constraints			
	AdjC	NcC	NmgC	VsbC
Precision [%]	97.39	67.78	92.36	80.36

Table 3.6: The accuracy of the lexico-morphosyntactic constraints

As one could expect, the highest accuracy was achieved for the AdjC constraint, based strongly on agreement. The tagger caused the majority of the errors. In some cases an adjective located between two nouns of the same values of the analysed grammatical categories was mistakenly associated with the wrong noun. The good result of NmgC was in large extent artificially increased by the aforementioned loose definition of the genitive nominal modifier assumed in NmgC and its evaluation. For example, we did not distinguish genitive arguments of a gerund which modifies the head from the proper genitive modifiers of the head. Still, it is worth noting that we have achieved relatively good results of subject identification using a fairly simple constraint mechanism VsbC.

As the majority constraints for verbal and adjectival LUs are symmetrical or very similar to those for nominal LUs, we expect similar accuracy.

3.4.4 Transformation based on rank weighting

In the co-incidence matrix constructed in step 2 (Section 3.4.2, p. 65) as a result of the general MSR extraction process, each LU is described by a vector of features that correspond to all context types taken into consideration. The initial value of features are the frequencies of occurrences of the given target LU in the corresponding lexico-syntactic contexts. This raw information, however, is very noisy and not reliable.

- Some features deliver little or no information. Consider, for example, very frequent adjectives with vague meaning, such as “nowy” (*new*, 627874 occurrences in corpora) or “wielki” (*large, great*, 615785), “mój” (*mine*, 592976), or very frequent verbs that occur with many subjects, such as “być” (*be*, 6944204), “mieć” (*have*, 2332773), and so on. They result in large values of the corresponding features (frequencies), occur with the majority of target LUs and make every LU related to every other LU.

- Accidental feature values caused by very infrequent, mostly singular, occurrences of the corresponding lexical elements with the target LUs have negligible influence on the well-described frequent target LUs with many non-zero features, but can relate some infrequent LUs to many others just because of a few accidental feature values, e.g. association of “pies” (*dog*) with “żelbeton” (*reinforced concrete*) found by noun-coordination constraint (NcC).

- Raw feature values can also be biased by corpora in two ways: values of features from some subset can be increased (e.g., some specific modifiers repeatedly used across some set of documents) and for some subset of the target LUs the average level of the values of their features can be increased in comparison to the rest of the target LUs (e.g., because LUs from the given subset occur more frequently in the corpora).

Thus, most MSR extraction methods *transform* the initial raw frequencies before the final computation of the MSR value. Such a transformation is typically a combination of *filtering* and *weighting*. The quality and behaviour of an MSR depend to a large extent on the transformation applied. For example, in (Piasecki et al., 2007a), the increase from 82.72% of accuracy in WBST+H to 86.09% was achieved only by changing the transformation.

Transformations proposed in the literature usually combine initial filtering based on simple heuristics referring to frequencies with weighting based on the analysis of statistical association between the given target LU and features. The filtering functions can be applied to both target LUs and features, in order to remove target LUs for which we do not have enough information, or to exclude from description features that do not deliver enough information. Mostly, a filtering function is defined as a simple comparison with the threshold. In the case of target LUs, LU frequency and the number of non-zero features are tested, e.g. Lin (1998) filtered out all target LUs occurring less than 100 times in the corpus of about 64 million words. For features, elimination criteria can be based on the number of LUs described (the number of non-zero cells for the given feature), total feature frequency (with all LUs: a sum over the column) or statistical analysis, as in (Geffet and Dagan, 2004) discussed in a while.

Hindle (1990) applied *Mutual Information* (MI) to compute feature weights in relation to particular target LUs. Weights express the strength of association between a target LU and a feature. Lin (1998) used a slightly modified version of MI based on the ratio of the information shared and total description. Lin and Pantel (2002) applied *Pointwise Mutual Information* modified by a discounting factor during LU semantic similarity computation – see the generalised version in (Mohammad and Hirst, 2006). Geffet and Dagan (2004) introduced *Relative Feature Focus* (RFF), a feature-weighting function based on a two-step transformation. First they extract Lin's MSR, filtering out features with the overall frequency below 10 and MI weight below 4. Next, they recompute the value of a feature f in LU u as the sum of MSR values of LUs most related to u such that they have a non-zero value for f. The final MSR(RFF) is calculated from the new feature values. Weeds and Weir (2005) proposed "a flexible, parameterised framework for calculating" MSR, based on the idea of casting the problem as a *Co-occurrence Retrieval Model* (CRM). CRM describes semantic relatedness of two LUs in terms of weighted precision and recall of feature sharing between them. Of several weighting functions applied, the best results came with MI and *t-score* measures.

The method of transformation is often tightly coupled with the computation of the final MSR value, e.g. (Lin, 1998, Weeds and Weir, 2005), but vector similarity measures independent of the weight function are also applied, e.g. cosine measures or Jaccard coefficient (Mohammad and Hirst, 2006). All these transformations still associate the calculated value of a feature with the initial frequency. For example, in the case of Lin's measure more frequent features do not only get values higher than less frequent features; the value level of frequent features is also higher than the value level of those less frequent. We have noticed that this phenomenon negatively affects the accuracy of MSR (Piasecki et al., 2007a). In a new weighting function proposed, the z-score measure was combined with a *Rank Weight Function* [RWF], a transformation from values to *ranks*.

The main idea behind RWF is to put more emphasis in the description of LU meaning on the identification of features most relevant to this LU. The calculation of the *exact* values of the strength of their association with the target LU is less important. We believe that these values are largely the artefact of the biased corpus frequencies, so one should not depend on them too strictly during row vector similarity calculation. The particular order of relevance of the features delivers clearer information. In RWF, the meaning of the given LU is described by an *ordered set* of relevant features, and the meaning of two LUs can be compared on the corresponding sequences of features ordered by relevance. In order to keep the correlation between relevance and feature value, each feature is assigned a rank: its position in the *reversed order*, so the most relevant feature receives the highest number equal to the number of features selected for the given LU. We will shortly present an MSR algorithm based on RWF. Because differences among the relevance values can be very small, and thus accidental, we use a partial order. Some features are assigned the same rank and the same value. The version of RWF based on partial order has been called *Generalised RWF* [GRWF] (Broda et al., 2009). Independent of the frequencies of subsequent LUs, feature values are natural numbers taken from the same limited subset. Examples of the GRWF application appear in Tables 3.7 and 3.8.

AdjC	frequency	Lin	$GRWF_{Lin}$
mieszkalny (*residential*)	6173	5.37	6
nowy (*new*)	1776	0.89	2
komunalny (*communal*)	1362	3.65	5
gospodarczy (*related to household*)	1170	1.8	3
stary (*old*)	1141	1.67	3
główny (*principal*)	968	1.56	3
szkolny (*school-related*)	651	2.73	4
wysoki (*tall, high*)	646	0.43	1
wielorodzinny (*multi-family*)	639	5.73	7
zakładowy (*related to factory/establishment*)	522	3.62	5

Table 3.7: GRWF applied to adjectival features for the noun "budynek" (*building*); frequency is measured in the joint corpus (Section 3.4.5), 'Lin' is the weight calculated by the Lin's algorithm (Lin, 1998) and $GRWF_{Lin}$ – the rank computed by the Generalised RWF function utilising Lin's measure. We show 10 most frequent features with their respective values of weight by Lin's measure and position in ranking created with GRWF. The maximal value of Lin's weight in the matrix is 6.55. The highest rank in $GRWF_{Lin}$ is 8

Not weighted	rank – $GRWF_{Lin}$
mieszkalny (*residential*)	średniowysoki (*medium-high*)
nowy (*new*)	apartamentowy (*related to apartments*)
komunalny (*communal*)	celniczy (*customs-related*)
gospodarczy (*related to household*)	czterokondygnacyjny (*four-storey*)
stary (*old*)	dwukondygnacyjny (*two-storey*)
główny (*principal*)	dwunastopiętrowy (*twelve-storey*)
szkolny (*school-related*)	dziesięciopiętrowy (*ten-storey*)
wysoki (*tall. high*)	dziewięciokondygnacyjny (*nine-storey*)
wielorodzinny (*multi-family*)	dziesięciopiętrowy (*ten-storey*)
zakładowy (*related to factory*)	luksusowy (*luxury*)

Table 3.8: Most important adjectival features for the word *budynek* 'building'

In tests carried out for Polish and English corpora, MSR based on GRWF (henceforth MSR$_{GRWF}$ (the index signals the weight method applied) outperformed MSR based on other measures discussed (Broda et al., 2009).

We now present the main line of the process of transformation based on the Generalised RWF. For the sake of clarity, we omitted several variants of particular steps. The full specification of the algorithms, its variants and parameters can be found in (Broda et al., 2009).

1. Let $\mathbf{M}$ be a co-incidence matrix, w_i – a LU, c_j – a feature, $\mathbf{M}[w_i, c_j]$ – the co-occurrence frequency of w_i together with c_j.

2. For the given w_i, we recalculate the weighted values of the corresponding cells, using a *weight function* f_w equal to the Lin's MI: $\forall_c\ \mathbf{M}[w_i, c] \leftarrow f_w(\mathbf{M}[w_i, c])$.

3. The subset $F_k[w_i] \leftarrow f_k(\mathbf{M}[w_i, \bullet])$ of k the most relevant features of the given row vector is selected as a LU description (other features are set to 0).

4. Features from the set $F_{sel}[w_i]$ are sorted in the ascending partial order on the weighted values: features with the same value occupy the same position in the ranking.

5. For each selected feature c_j a new value is calculated:
   $\mathbf{M}[w_i, c_j] = non\_zero(\mathbf{M}[w_i, \bullet])$ - $f_{por}(c_j)$, where

   - non_zero returns the number of non-zero features in the given row,
   - $f_{por}(c_j)$ calculates the position of c_j, starting from zero in the partial order ranking based on f_w – a natural strategy is followed, with subsequent positions numbered consecutively.

In earlier experiments, weight functions different than Lin's variant of MI were used, but the MSR$_{GRWF}$ based on Lin's MI appeared to be significantly better during experiments presented in (Broda et al., 2009). The value of the k parameter was set experimentally to 10000, but the number of the features comprising an LU description is usually smaller.

By setting the highest feature value to the number of relevant features that comprise the description of the given LU, infrequent or specific LUs described by only few features are differentiated from the well-described LUs with many features. MSRs calculated for LUs of these two groups are lower but closer to intuition than in the case of assigning the identical highest feature value across LUs.

3.4.5 Benefits for wordnet construction

Our aim in MSR extraction was ultimately to use MSR as the basic knowledge source in semi-automatically expanding plWordNet with new synsets and relation instances. We planned to obtain material for new synsets by clustering target LUs by MSR-produced values. We discuss the experiments next, in Section 3.5. We also hoped to find on the MSRlist$_{(x,k)}$ lists new instances of wordnet relations between LUs present in plWordNet and new LUs, as well among new LUs. We assumed manual verification of the results. An MSR was constructed for a set of one-word and two-word nominal LUs[16] including all LUs already present in the core plWordNet, as well a set of LUs planned as the basis for expansion. The drawbacks of a pure corpus approach, discussed in Section 2.4, made us take a more dictionary-based approach in defining the lemma list for the expansion of plWordNet. In the end, 13285 nominal LUs have been selected for extracting an MSR for nominals:

- 5340 nominal lemmas described in the core plWordNet,
- additional lemmas (further on referred to as *new lemmas*):
  - nominal lemmas acquired from a small Polish-English dictionary (Piotrowski and Saloni, 1999),
  - two-word LUs from a general dictionary of Polish (PWN, 2007),
  - the lemmas that occur over 1000 times in the largest available corpus of Polish, IPIC (Przepiórkowski, 2004),

The small Polish-English dictionary (Piotrowski and Saloni, 1999) was used as the main source, because its small size makes its entries close to the core of the Polish vocabulary.

First experiments on MSR extraction had been performed only on IPIC (Piasecki and Broda, 2007, Piasecki et al., 2007a,b). Later, when we collected other corpora, we observed a correlation between an increase in WBST+H and the increasing size of the overall corpus used. The final version of the nominal MSR has been extracted from three corpora:

- IPIC (including about 254 million tokens) (Przepiórkowski, 2004) (it is not balanced but it covers a variety of genres: literature, poetry, newspapers, legal texts, stenographic parliamentary records and scientific texts);
- a corpus of the electronic edition of a Polish newspaper *Rzeczpospolita* from January 1993 to March 2002 (about 113 million tokens) (Rzeczpospolita, 2008);

[16] We were limited to at most two-word LUs by the technology of the extraction of multiword expression we had developed (Broda et al., 2008).

- and a corpus of large texts in Polish (about 214 million tokens) collected from the Internet; only documents containing a small percentage of erroneous word forms (tested manually) and not duplicated in the other two corpora were included in the collected corpus.

Henceforth, we will refer to all three corpora used together as the *joint corpus*.

For nominal MSR, fours types of lexico-morphosyntactic constraints have been used (Section 3.4.3): `AdjC`, `NcC`, `NmgC` and `VsbC`. Lists of lexical elements have been defined as the combination of one-word LUs in the joint corpus and two-word LUs assumed for the expansion of plWordNet. We also used 63328 adjectives and adjectival participles for `AdjC`, 199250 one-word and two-word nominal LUs for `NcC` and `NmgC`, and 29564 verbs for `VsbC`.

Evaluation of the extracted nominal MSRs was performed on the basis of the WBST+H and EWBST tests presented in Section 3.3. The WBST+H test used for the final version of the nominal MSR, generated from the plWordNet version 11.2008 consisted of 9486 questions; EWBST had 8689. Table 3.9 shows the results of the WBST+H and EWBST. Table 3.10 includes the results in relation to particular types of constraints (we constructed several coincidence matrices, from which different MSRs were built and tested).

all LUs		*more frequent than* $\geq$ 1000	
WBST+H	EWBST	WBST+H	EWBST
88.14	69.75	92.28	75.43

Table 3.9: The accuracy of the nominal MSR based on the generalised RWF and Lin's version of MI

`AdjC`		`NcC`		`NmgC`		`VsbC`		*all*	
$\geq 10^3$	*all*	$\geq 10^3$	*all*	$\geq 10^3$	*all*	$\geq 10^3$	*all*	$\geq 10^3$	*all*
90.90	84.98	88.81	80.67	76.25	65.10	79.17	65.89	**92.28**	**88.14**

Table 3.10: The accuracy [%] of nominal MSRs based on different morphosyntactic constraints; all MSRs use Generalised RWF based on Lin's MI. "$\geq 10^3$" means more frequent than 1000

The best results achieved for the nominal MSR in WBST+H and EWBST (Table 3.9) are close to the average human results: 86.64% and 71.34%, respectively (Section 3.3). Both tests can clearly be interpreted from the perspective of practical application of the MSR. It can distinguish among semantically related and unrelated LUs with the accuracy 88.14% (WBST+H) and semantically closely related and more remotely related with accuracy 69.75% (EWBST). Moreover, our comparison of the MSR_{GRWF} with several other MSRs based on methods proposed in the literature

(Broda et al., 2009) showed that one can hardly expect to achieve a significantly better result with any other MSR.

A closer inspection of MSRlist$_{(x,k)}$ lists – Tables 3.11 and 3.12 – shows examples – reveals, however, that (though many pairs are clearly semantically related) the percentage of instances of wordnet relations is much below the psychological barrier of 50%. It is also very hard to find any clear threshold above which the MSR value guarantees that a given pair of LUs is a instance of a wordnet relation[17]. These intuitions were confirmed in an experiment with the manual analysis of the 364 LU pairs from MSRlist$_{(x,k)}$ lists. The pairs were selected randomly from the MSR$_{RWF}$ extracted from IPIC for the needs of (Derwojedowa et al., 2008). There was a manual assessment of each pair $\langle x, y\rangle$ such that $y \in MSRlist_{(x,k)}$ and $MSR(x, y) \geq \tau_{MSR}$[18] as belonging to one of the wordnet relations. Half of the pairs did not belong to 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 hyponyms or hypernyms (not necessarily direct) already present in plWordNet, and 16% of new close hyponyms/hypernyms and co-hyponyms were discovered. 1% of known meronyms and holonyms were found and 5% of new ones were discovered.

The size of the corpus used for MSR extraction is significant for MSR's accuracy. We therefore repeated in 2008 the experiment with manual evaluation of the MSRlist$_{(x,k)}$ list. We used an MSR$_{RWF}$ extracted from the joint corpus for the final plWordNet expansion assisted by the WordNet Weaver system, see Section 4.5. 1000 LU pairs were randomly selected and evaluated by one of the co-authors, and assigned to one of the classes described below:

- 523 LU pairs were not instances of any wordnet relation, 4 LU pairs included errors caused by the morphosyntactic preprocessing (such as a non-word form generated by the morphological guesser), and 5 pairs contained an incomplete multiword LU, but in the hypernymy relation to the second member of the pair,
- 228 LU pairs included elements linked by the synonymy or hypernymy/hyponymy relation (the latter not necessarily direct) — only 16 instances had been already described in plWordNet,
- 158 LU pairs were co-hyponyms or close "cousins" (indirect co-hyponyms),
- 66 LU pairs represented meronymy/holonymy, and 12 pairs were co-meronyms,
- 4 LU pairs were antonyms.

[17] The MSR values seem not to be directly comparable among different target LUs for which MSRlist$_{(x,k)}$ lists are generated.

[18] We set τ_{MSR} to 0.2.

The comparison of both evaluations performed on two different development versions of MSR$_{RWF}$ shows that instead of the increasing accuracy of the measure in the WBST+H test, the percentage of wordnet relation instances remains stable. We need additional extraction mechanisms in order to increase the percentage of the target instances in the results and differentiate between wordnet relations – see Chapter 4.

According to the planned semi-automatic expansion of the adjective and verb parts of plWordNet, the respective MSRs were extracted using the joint corpus and the MSR$_{RWF}$ and MSR$_{GRWF}$ algorithms. The procedures followed the blueprint adopted for the nominal MSR. We acquired two sets, 4668 adjectival lemmas and 17990 verbal lemmas[19]. They came from the core plWordNet (2618 and 3239, respectively), the small Polish-English dictionary (Piotrowski and Saloni, 1999) and the joint corpus (those occurring $\geq$ 1000 times).

Both MSRs were tested with WBST+H tests including 2814 QA pairs for adjectival lemmas and 5484 for verbal lemmas. The QA pairs encompass 1574 different adjectival lemmas (among them 959 occur over 1000 times in the joint corpus) and 2960 different verbal lemmas (1902 occur more than 1000 times). Some of them occur in QA pairs more than once but with different near-synonyms.

Tables 3.13 and 3.14 show the results for different MSRs on the same tests for LUs of different frequency. For WBST+H the baseline random selection is 25%. We divided the analysed adjectival and verbal lemmas into two groups by their frequency in IPIC: those occurring $>$ 1000 and the others. The results for the first group are given in Table 3.13. In Table 3.14 we present results obtained for all LUs.

Working with the same generated co-incidence matrices for verbs and adjectives, we compared the application of RWF with three other measures: Lin's measure (Lin, 1998), CRMI (Weeds and Weir, 2005), RFF (Geffet and Dagan, 2004). From a large number of proposed solutions, we selected only the measures based on lexico-syntactic features. Lin's measure was included in the set because of its significant influence on the subsequent research. CRMI has been extensively compared with several other approaches showing significant improvement. RFF was chosen for the idea of feature selection present in it. RFF is calculated in two phases: in the first phase features are evaluated and the best 100 are selected, re-weighted and used in LU similarity calculation in the second phase. In all three approaches the similarity computation is based in some way on Mutual Information weighting, which is also often used by other methods. Finally, the approach of Freitag et al. (2005) is one of the few that deal with the similarity of adjectives and verbs.

In the case of RWF, we also determined experimentally the threshold k for the number of features selected achieving the best results with

[19] Besides one-word lemmas, we only considered verbs paired with the reflexive particle *się*.

język polski (*Polish language*) (G)	
język angielski (*English language*)	0.218
język niemiecki (*German language*)	0.207
język francuski (*French language*)	0.177
język rosyjski (*Russian language*)	0.175
język czeski (*Czech language*)	0.168
język węgierski (*Hungarian language*)	0.163
język słowacki (*Slovak language*)	0.140
matematyka (*mathematics*)	0.139
geografia (*geography*)	0.137
filologia (*philology*)	0.137
język łaciński (*Latin language*)	0.131
greka (*Greek language*)	0.127
językoznawstwo (*linguistics*)	0.127
język grecki (*Greek language*)	0.123
język literacki (*literary language*)	0.122
gramatyka (*grammar*)	0.120
język hebrajski (*Hebrew language*)	0.115
łacina (*Latin language*)	0.112
informatyka (*informatics*)	0.111
polszczyzna (*Polish language*)	0.110

gaz ziemny (*natural gas*) (G)	
gaz (*gas*)	0.258
węgiel kamienny (*coal (pit-coal)*)	0.207
węgiel brunatny (*brown coal*)	0.197
ropa (*oil*)	0.193
olej opałowy (*heating oil*)	0.164
paliwo (*fuel*)	0.161
wodór (*hydrogen*)	0.160
kopalina (*fossil (mineral)*)	0.160
węgiel (*coal*)	0.143
olej napędowy (*diesel fuel*)	0.140
gaz płynny (*liquid gas*)	0.140
koks (*cox*)	0.127
ołów (*lead*)	0.119
azot (*nitrogen*)	0.119
tlen (*oxygen*)	0.116
uran (*uranium*)	0.116
biokomponent (*biocomponent*)	0.115
cynk (*zinc*)	0.114
łupek palny (*slate (fuel)*)	0.113
benzyna (*gasoline*)	0.110

Table 3.11: Examples of lists (G means *good*) of the 20 LUs most similar to the given one according to the MSR based on RWF(z-score)

jacht (*yacht*) (G)	
motorówka (*speedboat*)	0.268
łódź (*boat*)	0.234
statek (*ship*)	0.233
prom (*ferry*)	0.202
łódka (*boat*)	0.196
żaglówka (*sailboat*)	0.190
kuter (*cutter*)	0.189
okręt (*liner ship*)	0.186
fregata (*frigate*)	0.175
frachtowiec (*cargo ship*)	0.175
tankowiec (*tanker*)	0.171
szalupa (*launch*)	0.160
holownik (*tugboat*)	0.159
szkuner (*schooner*)	0.157
samolot (*airplane*)	0.157
barka (*barge*)	0.156
ponton (*pontoon*)	0.156
tratwa (*raft*)	0.153
wodolot (*hydroplane*)	0.149
kajak (*kayak*)	0.141

kamieniołom (*quarry*) (Acc)	
kopalnia (*mine*)	0.136
ogródek działkowy (*allotment*)	0.107
gorzelnia (*distillery*)	0.092
złoże (*lode*)	0.088
wysypisko (*dump*)	0.087
wyrobisko (*excavation*)	0.086
ogród botaniczny (*botanical garden*)	0.079
żwirowisko (*gravel pit*)	0.078
hotel robotniczy (*(employee) hostel*)	0.078
orlik grubodzioby (*eagle (Aquila clanga)*)	0.075
oczyszczalnia (*purification plant*)	0.074
folwark (*grange*)	0.073
pieczara (*cave*)	0.072
karczma (*inn*)	0.072
rów melioracyjny (*drainage ditch*)	0.070
składowisko (*storage yard*)	0.069
żłobek (*creche*)	0.069
uzdrowisko (*spa*)	0.069
kąpielisko (*resort*)	0.068
pensjonat (*pension*)	0.068

Table 3.12: Examples of lists (G – *good* and Acc – *accidental*) of the 20 LUs most similar to the given one according to the MSR based on RWF(z-score)

Features	Lin	CRM_{MI}	PMI	RWF_{zscore}	$GRWF_{Lin}$
NArg(acc)	69.17	57.52	63.68	62.51	**70.86**
NArg(dat)	50.00	24.54	46.16	28.19	**50.10**
NArg(inst)	65.37	51.52	58.32	46.44	**67.97**
NArg(loc)	63.02	54.67	57.71	47.78	**65.81**
Nsb	63.68	56.41	57.65	**66.32**	65.59
VPart	55.81	51.30	53.11	54.10	**56.70**
VAdv	75.21	60.06	64.00	72.44	**75.49**
NArg(acc+dat+inst+loc)	72.03	64.57	68.95	68.70	**73.87**
NSb+NArg+VPart+VAdv	74.16	64.83	70.86	**75.94**	75.33
AAdv	66.15	21.57	58.77	63.86	**67.02**
AA	80.41	72.16	77.25	74.95	**81.90**
ANmod	81.96	75.39	80.60	**83.57**	82.46
ANmod+AAdv	82.33	75.08	81.34	**85.00**	83.26
ANmod+AA	82.77	76.94	83.70	**86.42**	83.39
ANmod+AAdv+AA	84.44	76.63	83.70	**86.92**	86.55

Table 3.13: Experiments with MSRs for frequent lemmas ($>$ 1000 occurrences in joined corpora)

- $k = 10000$ for the frequent adjectives, $k = 1000$ for the frequent verbs
- $k = 1000$ for all adjectives, and $k = 1000$ for all verbs.

An automatic mechanism of the k value adjustment on the basis of data analysis would be a valuable extension of the RWF method. It must be noted, however, that the range of results achieved for different k values is limited. For example, in the case of frequent verbs and the joint matrix NSb+NArg+VPart+VAdv we get 73.23% for $k = 100$, 76.24% for $k = 500$, **77.12%** for $k = 1000$ and 76.88% $k = 5000$. Results become stable around $k = 300$ and only a slight tuning is required by finding the optimal value of k. There was a similar result for nouns (Piasecki et al., 2007b).

In the case of verb constraints, the highest results by a single type of a constraint is generated, surprisingly, by a simple closest adverb identification. NArg(dat) and NArg(inst) matrices are too sparse and the identification of a subject generates too many errors (we do not apply any parser). For a joined matrix, however, RWF selects features effectively enough to achieve a result that is significantly better than any single verb matrix.

In the case of adjectives, the differences of accuracy achieved for different types of constraints are much smaller. The joined matrix is also better than any single one. Hatzivassiloglou and McKeown (1993) claim that co-occurrence of two adjectives in one noun phrase (clearly indicated in Polish by their morphosyntactic agreement) is a negative feature. This claim is contradicted by the result of AA alone and AA combined with other matrices.

Features	Lin	CRMI	PMI	RWF_{zscore}	$GRWF_{Lin}$
NArg(acc)	60.47	58.13	55.53	56.89	**63.35**
NArg(dat)	38.59	23.98	36.40	26.29	**39.06**
NArg(inst)	55.60	45.59	50.75	42.54	**57.57**
NArg(loc)	51.42	46.72	48.54	41.47	**54.50**
Nsb	53.15	54.54	47.68	**57.79**	54.78
VPart	46.70	44.69	44.89	47.28	**48.32**
VAdv	65.32	53.77	58.04	64.19	**66.50**
NArg(acc+dat+inst+loc)	65.13	65.04	60.94	64.17	**68.05**
NSb+NArg+VPart+VAdv	67.10	67.80	62.42	62.41	**71.85**
AAdv	57.60	20.86	53.27	**58.96**	58.71
AA	74.24	71.86	71.75	72.32	**76.87**
ANmod	76.12	74.77	73.99	**79.18**	77.75
ANmod+AAdv	76.97	75.41	74.80	**81.13**	78.93
ANmod+AA	78.18	78.32	78.43	**83.05**	79.89
ANmod+AAdv+AA	79.71	78.32	78.39	**83.26**	82.48

Table 3.14: Experiments with MSRs for all lemmas

The result of our best adjective MSR is very close to the result achieved by humans (Section 3.3.1). For verbs, the difference is comparable to that observed for nouns (Piasecki et al., 2007b) (but the result of verb MSR still approaches human performance).

The constructed MSRs are intended to assist linguists in selecting LUs semantically related to the LU being edited. Lexicographers can find missing synonyms or instances of lexico-semantic relations while browsing the MSRlist$_{(x,k)}$ lists (according to the MSRs).

Long suggestion lists may preclude careful analysis. We chose $k = 20$ for a small experiment to test a possible future use of both MSRs by linguists. We randomly selected two subsets of lemmas, verbs and adjectives. We determined sample sizes in such a way that the results of the manual evaluation performed on the samples could be ascribed to the whole sets with the 95% confidence level, according to the method discussed in (Israel, 1992). For every LU in each subset, we generated the list of the $k = 20$ LUs most related to the given one. One of the co-authors manually assessed all elements on all lists, distinguishing any elements that are in some wordnet relation to the head LU.

The evaluated LU lists were classified into:

- *very useful* – a half, or almost a half, of the LUs on the list are in some semantic relation to the given one,
- *useful* – a sizable part of the list is somehow related,

- *neutral* – several LUs on the list are in some relation, but the linguist might miss them,
- *useless* – at most a few LUs may be related.

The results of the manual evaluation appear in Table 3.15.

PoS	very useful	useful	neutral	useless	no relations
Verb [%]	17.8	37.6	20.0	15.6	9.0
Adjective [%]	19.2	26.3	29.7	14.4	10.4

Table 3.15: Manual evaluation of MSR for verbs and adjectives performed for (Broda et al., 2008) on the MSRs from that time

Selected lists for verbs and adjectives are shown in Tab. 3.16. The English translations “select” the meaning common to the grouping that the list suggests.

In nearly half of the cases, the linguist can find valuable hints on the list generated on the basis of MSRs. Suggestions should help notice specific or domain-restricted uses of LUs. The manual evaluation suggests MSR accuracy much lower than for the WBST, but the latter operates on generic semantic similarity rather than specific semantic relations. However, besides the relatively large percentage of the MSRlist$_{(x,k)}$ lists evaluated as very useful and useful, the percentage of correct hints – pairs of LUs in some wordnet relation – was still significantly below the 50% threshold of acceptance for the MSRlists$_{(x,k)}$ as a valuable tool for linguists. In order to increase the accuracy of automatic extraction of instances of wordnet relations, we need to look for additional knowledge sources. In the case of adjectival LUs and verbal LUs the use of lexico-syntactic patterns is hardly possible, cf Section 4. Only antonymy for adjectives seems to be marked by specific language expressions. If we redefine the support task to expanding a wordnet, the wordnet structure already in place becomes an additional source of knowledge. We will explore this approach in relation to verbal LUs in experiments presented in Section 4.5.3.

In the case of all three parts of speech, the constructed MSRs have the accuracy in WBST+H and EWBST (nominal MSR only) which surpasses that of the MSRs based on the algorithms proposed in literature. Due to the nature of the applied tests, we can conclude that all three MRS extracted have the ability to distinguish among semantically related and unrelated LUs with an accuracy that is relatively close to the average results of humans in the same task. Let us note that the results may have been different had the users been trained linguists. Nevertheless, having these three good MSRs, we have still not achieved a tool of practical importance for semi-automatically expanding the core plWordNet: generated MSRlist$_{(x,k)}$ lists include too many LU pairs which do not belong to any wordnet relation. Our experience makes us pessimistic about the

strzec (*guard, protect*) (`G`)		**chromować (*chrome*)** (`Acc`)	
pilnować (*guard*)	0.141	niklować (*nickel*)	0.145
patrolować (*patrol*)	0.117	mocować (*fasten*)	0.110
ufortyfikować (*fortify*$_{perf}$)	0.104	zamocować (*fasten*$_{perf}$)	0.110
chronić (*protect*)	0.087	skorodować (*corrode*$_{perf}$)	0.109
ochronić (*protect*$_{perf}$)	0.079	przymocować (*fasten*$_{perf}$)	0.108
zabezpieczyć (*secure*$_{perf}$)	0.076	wypolerować (*polish*$_{perf}$)	0.108
otoczyć (*surround*$_{perf}$)	0.075	umocować (*fasten*$_{perf}$)	0.107
czuwać (*watch*)	0.070	powyginać (*bend all over*$_{perf}$)	0.107
bronić (*defend*)	0.069	ocynkować (*zinc*$_{perf}$)	0.107
zaminować (*mine*$_{perf}$)	0.069	rzeźbić (*carve*)	0.107
zrujnować (*ruin*$_{perf}$)	0.067	obluzować (*loosen up*$_{perf}$)	0.106
usytuować (*situate*$_{perf}$)	0.067	pogiąć (*bend*$_{perf}$)	0.106
eskortować (*escort*)	0.062	przytwierdzić (*attach*$_{perf}$)	0.101
oświetlić (*light*$_{perf}$)	0.061	wygiąć (*bend*)	0.099
zagradzać (*fence*)	0.060	przyśrubować (*screw down*$_{perf}$)	0.094
ogrodzić (*fence*$_{perf}$)	0.060	wystawić (*exhibit, put out*)	0.090
zaryglować (*bolt*$_{perf}$)	0.059	błyszczeć (*shine*)	0.088
ostrzeliwać (*bombard*)	0.059	złocić (*gild*)	0.088
oznakować (*mark*$_{perf}$)	0.058	zamontować (*fit onto*$_{perf}$)	0.085
stacjonować (*station*)	0.057	dokręcić (*fasten (a screw)*$_{perf}$)	0.083

ognisty (*fiery*) (`G`)		**czołowy (*leading*)** (`Acc`)	
świetlisty (*shiny*)	0.295	wybitny (*prominent*)	0.336
płomienny (*blazing*)	0.273	znany (*known*)	0.318
srebrzysty (*silvery*)	0.230	znakomity (*illustrious*)	0.301
złocisty (*golden*)	0.229	utalentowany (*talented*)	0.273
płomienisty (*flaming*)	0.225	amerykański (*American*)	0.247
szkarłatny (*scarlet*)	0.219	doświadczony (*experienced*)	0.244
świetlny (*light-related*)	0.215	polski (*Polish*)	0.242
czarny (*black*)	0.204	świetny (*excellent*)	0.240
czerwony (*red*)	0.197	dobry (*good*)	0.235
purpurowy (*crimson*)	0.192	francuski (*French*)	0.235
pomarańczowy (*orange*)	0.191	szwedzki (*Swedish*)	0.231
błękitny (*light blue*)	0.190	włoski (*Italian*)	0.224
migotliwy (*shimmering*)	0.186	słynny (*famous*)	0.223
niewidzialny (*invisible*)	0.183	brytyjski (*British*)	0.219
lodowy (*icy*)	0.179	austriacki (*Austrian*)	0.218
śmiercionośny (*lethal*)	0.177	czeski (*Czech*)	0.208
krwawy (*bloody*)	0.176	fiński (*Finnish*)	0.207
złoty (*gold*)	0.175	rosyjski (*Russian*)	0.207
tęczowy (*rainbow*)	0.174	niemiecki (*German*)	0.204
biały (*white*)	0.173	młody (*young*)	0.201

Table 3.16: Examples of lists (`G` — *good* and `Acc` — *accidental*) of the 20 LUs most similar to the given one for verbs and adjectives

possibility of constructing an MSR significantly better in this respect. Semantic and pragmatic constraints make many LUs semantically related to many other LUs and MRSs based on distributional semantics generate a continuum of relatedness values for pairs of LUs. Wordnet relations appear as just weakly identifiable characteristic subspaces in the continuum of semantic relatedness. We need an additional way of selecting those LU pairs from the MSRlist$_{(x,k)}$ lists which represent particular wordnet relations. Two ways appear to emerge:

1. application of lexico-syntactic patterns (Sections 3.2 and 4) as an additional source of knowledge,

2. introduction of an additional classifier trained on the plWordNet data and used for filtering out MSRlist$_{(x,k)}$ pairs which are not instances of any wordnet relation (Section 4.5.1).

We also mentioned briefly in the discussion of verbal and adjectival MSRs the idea of changing the perspective from automatic extraction of sets of instances of the wordnet relations to expanding the existing wordnet with new lemmas anchored in existing synsets. This can significantly extend the amount of knowledge available and reduce the complexity of the problem. We will present in Section 4.5.3 a solution following this idea. Let us emphasise here that it is automated wordnet expansion which was our assumed goal, not automatic wordnet construction from scratch.

3.5 Sense Discovery by Clustering

The synset is one of the most fundamental building blocks of the wordnet structure. An algorithm for automatic extraction of synsets would be very helpful for linguists who build a wordnet up manually (though usually with substantial software support). *Clustering* groups objects on a hyperplane so as to minimise the distance between objects inside a group and maximise the distance between objects from different groups. A definition of distance, or similarity, between objects is required for such grouping. For clustering of lemmas into synset-like groups, we could use directly a Measure of Semantic Relatedness [MSR] (Section 3.4). A drawback would be that MSRs tend to merge different lemma senses in one vector that represents the meaning of lemmas, or to over-represent one predominant sense of a given lemma (Piasecki et al., 2007a). That is why we need a clustering method aware of ambiguity in lemma meaning.

The *Most Frequent Sense* heuristic states that in one genre or domain one sense of a given lemma is dominant (Agirre and Edmonds, 2006). Without thematically labelled corpora one can hope that clustering techniques make it possible to achieve approximation of domains, because documents are grouped by similarity. On the other hand, the *one-sense-per-discourse* heuristic states that a lemma is used only in one of its senses in one discourse (Agirre and Edmonds, 2006). Combining both heuristics, we can assume that polysemous lemmas will be used in one dominant sense in one cluster.

One can hope to alleviate the ambiguity present in MSR by incorporating knowledge of the domain of the documents that contain the given lemma. Document hierarchy could also be used as a base structure for a wordnet (or only parts of a wordnet).

The approach based on document clustering in sense discovery is discussed in Section 3.5.1. Another remedy for inherent polysemy in any MSR can be a specialized algorithm, for example *Clustering by Committee* [CBC] (Pantel, 2003). Section 3.5.3 describes an adaptation of CBC to Polish, and an extension.

3.5.1 Document clustering in sense discovery

Document clustering in our work had two reasons. First, we wanted to explore the possibilities of extracting knowledge about polysemy of lemmas from document groups. One-sense-per-discourse heuristic suggests that a polysemous lemma will appear in a given domain only in one of its meanings. On the other hand, document clusters can be labelled with keywords – most representative words for a document group. Arranging document clusters in a hierarchical tree could form the basic structure for a wordnet.

There are many clustering algorithms. Following a review of the possibilities (Jain et al., 1999, Forster, 2006, Broda, 2007) we chose two algorithms for further analysis and experiments. We looked at following properties of clustering algorithms: the ability to cluster high-dimensional data (such as documents represented by vectors), the ability to detect clusters of irregular shapes and the possibility of building hierarchical trees.

There are many ways of representing documents for clustering (Forster, 2006, Broda, 2007). In this work we used the Vector Space Model. In this model documents are represented as vectors in high-dimensional space. Each dimension of the space corresponds to occurrences of a specific word. Vectors store data describing occurrences of words in documents.

RObust Clustering using linKs [ROCK] (Guha et al., 2000) follows the agglomerative clustering scheme. Initially, each document is in a one-element cluster. Pairs of the most similar clusters are merged iteratively. The algorithm differs from others in how the merging is decided. ROCK selects for merging a cluster that maximises the number of *links* between documents. To avoid oversized clusters (or even putting all documents into one cluster), the algorithms imposes an expected number of links for a cluster of a given size.

The notion of links can be explained by *common neighbours*. Neighbourhood is defined using a similarity function: if two documents are similar enough, they are considered neighbours. If links replace similarity in clustering, global information about documents can be used; for details see (Guha et al., 2000) and (Broda and Piasecki, 2008a). Even if documents are not very similar, they can form a cluster in ROCK if they have many common neighbours. This clusters of unusual shapes possible.

The other clustering algorithm we considered is called *Growing Hierarchical Self-Organising Map* [GHSOM] (Rauber et al., 2002) an extension of Self-Organising Map [SOM] (Kohonen et al., 2000). SOM is an *artificial neural network*. Every neuron consists of a weight vector and a vector of positions in the map[20]. Training SOM is done in an unsupervised manner by applying a *winner-takes-all* strategy. Every document is delivered to the network several times. The neuron most similar to a given document is the *winner*. Weights of the winning neuron and neurons in its neighbourhood[21] are updated to be even more similar to the input pattern. The learning algorithm is constructed so that the neighbourhood of a neuron and the rate of weight updating decrease over time.

The GHSOM algorithm addresses one of SOM's most important drawbacks – the *a priori* definition of the map structure. Rauber et al. (2002) proposed an algorithm for growing SOM both in terms of the number of map neurons and the hierarchy.

Clustering results will be used in the extraction of polysemy information, labelling clusters with keywords and generation of a basic structure for a wordnet, so we wanted to be sure to select clustering algorithms that performs well on collections of Polish documents.

There exists a few approaches to the evaluation of clustering (Forster, 2006). For example, one can study the theoretical properties of an algorithm, or measure some mathematical properties of the resulting clusters. In some domains those methods can be appropriate, but we argue that for the domain of documents the most suitable evaluation method is by referring to *external criteria*, such as a comparison of the results with manually created pre-existing categories.

Our evaluation used parts of the Polish daily paper "Dziennik Polski", included in the IPI PAN Corpus [IPIC] (Przepiórkowski, 2004). It has been manually partitioned into categories: *Economy*, *Sport*, *Magazine*, *Home News*, and so on. Both ROCK and GHSOM gave results satisfactory in comparison to the "Dziennik Polski" data (Broda and Piasecki, 2008a). A manual inspection of the produced clusters confirmed those results. We did not find any mixing of major topics in groups, for example there was no document from *Sport* put into clusters with documents talking about *Economy*. The algorithms also found more categories than are actually present in the corpus. For example, different sport disciplines were partitioned into separate groups. An important

[20] For us, a map is a two-dimensional grid, with neurons placed in the nodes of the grid. This is not the only possible representation for SOM: a map can be hexagon-based or neurons can be placed in a three dimensional space.

[21] Note that this neighbourhood is different from neighbourhood in ROCK. In SOM it is defined simply as certain number of neurons in the map that are close to the winning neuron.

drawback of ROCK is that it sometimes produces a very deep and unbalanced hierarchy. On the other hand, GHSOM assigned pairs of documents into one cluster which did not appear together in any manually created category more often then ROCK.

We wanted to label with *representative words* document groups clustered in a hierarchical tree. Words describing groups of documents closer to the root of the tree should be more general than words used for the documents in the leaves. Ideally, we would obtain a basic hypernymy structure for plWordNet (or at least instances of *is-a* relation) out of the assigned labels.

Keyword extraction can be supervised or unsupervised. Supervised algorithm requires ample manually constructed resources. We applied only such unsupervised methods that try to capture statistical properties of words occurrences to identify words which best describe the given document. The statistics can be counted locally, using data from a single document only, or estimated from a large body of text. To benefit from both local and global strategies, we extended the method proposed by Indyka-Piasecka (2004) with the algorithm of Matsuo and Ishizuka (2004) into a hybrid keyword extraction method.

Indyka-Piasecka (2004) assigns a weight w to every lemma l that occurs in each document of a group. Additionally, lemmas are filtered on the basis of their *document frequency* df_l, that is a number of documents in which lemma l occurred. Both rare and frequent lemmas are not good discriminator of document content (cf. Indyka-Piasecka, 2004). The weight w is calculated by using two weighting schemes:

$$tf.idf_{l,d} = tf_{l,d} \times \log \frac{N}{df_l} \tag{3.6}$$

and *cue validity*

$$cv = \frac{tf_{group}}{tf} \tag{3.7}$$

where *tf* and *df* denote term frequency and document frequency.

Matsuo and Ishizuka (2004) used a three–step–process to assign a weight to every lemma. First, all words in a document are reduced to their lemmas (basic morphological forms) and filtered on the basis of term frequency and a stoplist. Then, they cluster lemmas in a document using two algorithms. If two lemmas have similar distributions, it means that they belong to the same group. As a measure of the probability distribution similarity between two lemmas l_1 and l_2 (Matsuo and Ishizuka, 2004) used the Jensen–Shannon divergence[22]. Lemmas are also clustered when they follow similar co-occurrence pattern with other lemmas. This can be measured using *Mutual Information*.

[22] Jensen–Shannon divergence is a symmetrised and smoothed version of Kullback–Leibler divergence (Manning and Schütze, 2001).

After the creation of clusters, the weight w is assigned using χ^2 test for testing if there is a bias in occurrences of the lemma with the group. Within our approach, a weight for a lemma is calculated in a way combining the methods of Indyka-Piasecka (2004) and Matsuo and Ishizuka (2004):

$$w_l = \alpha \cdot \min_{tf.idf_l} + \beta \cdot cv_l + \gamma \cdot \chi_l^2, \tag{3.8}$$

where $\min_{tf.idf_l}$ is the minimal $tf.idf$ weight for the given term l across the documents in a cluster, α, β, γ are parameters controlling impact of every measure on final weight.

Words which are assigned the highest weights are used as labels for the group of documents in the cluster tree.

3.5.2 Benefits of document clusters for constructing a wordnet

Our ultimate goal in document clustering was to obtain the basic structure for plWordNet. Document group labels could be used as synsets and cluster tree as a hypernymy hierarchy.

We evaluated our approach on plWordNet. The automatically created thesaurus was compared with the plWordNet hypernymy hierarchy. This failed: only 86 hypernymic instances (word pairs) were present in the thesaurus, fewer than 1% of all relations. Clustering whole documents might be a reason of low accuracy, but experiments with document segmentation decreased the quality of clustering (Broda, 2007, Broda and Piasecki, 2008a). On the other hand, keyword extraction methods developed primarily for information retrieval are not suitable for the discovery of relations between words that describe different groups of documents.

The extracted group labels are still quite very descriptive. For example, a group of documents about "interventionist purchase of grain and harvest in the area of Małopolska" are labelled with *zboże* (*grain*), *pszenica* (*wheat*), *tona* (*tonne*), *rolnik* (*farmer*) and *agencja* (*agency*). Another possible use of extracted words is to measure the degree of polysemy, because different meanings of words occurs in different branches of hierarchy.

3.5.3 Clustering by Committee as an example of word sense discovery

A good MSR can provide valuable information about word similarity during wordnet construction. For every word x, an MSR can produce a list of its k most similar words (denoted as $MSRlist(x, k)$) . Because of the nature of MSRs, those lists consists not only of words related by one lexico-semantic relation (Section 3.4). Part of the words on those *similarity lists* can be even unrelated to the *target word*. Choosing the right value for k can also be problematic. Not only does it depend on the MSR algorithm, but also the training phase can influence it. Worse still, the value of “good” k can change with word x for the same MSR.

Clustering techniques may help create better lists or groups of words. We would like to find a method that identifies lists of tightly interlinked word groups representing near-synonymy and close hypernymy, which could be added to plWordNet with as little intervention of the linguists as possible.

Standard partitioning clustering methods are ill-suited to the task of clustering lemmas. They can assign one word to a single cluster, which is problematic for polysemous lemmas. For lemmas that have one predominant meaning, only a cluster for one sense will be created. For polysemous lemmas without a predominant meaning the situation may be even less pleasant: such lemmas can lead to the creation of clusters that mix lemmas that have more than one of the polysemous lemma senses. That is why we need specialized clustering method.

Several clustering algorithms for the task of grouping words have been discussed in the literature. Among them, Clustering by Committee [CBC] (Pantel, 2003, Lin and Pantel, 2002) has been reported to achieve especially good accuracy with respect to evaluation performed on the basis of PWN. It is often referred to in the literature as one of the most interesting clustering algorithms (Pedersen, 2006).

CBC relies only on a modestly advanced dependency parser and on a MSR based on Pointwise Mutual Information [PMI] extended with a discounting factor (Lin and Pantel, 2002). This MSR is a modification of Lin’s measure (Lin, 1998) analysed in Section 3.4 and in (Broda et al., 2008) in application to Polish. Both measures are close to the RWF measure (Piasecki et al., 2007a) that achieves good accuracy in synonymy tests generated out of plWordNet (Section 3.3).

Applications of CBC to languages other than English are rarely reported in the literature. Tomuro et al. (2007) mentioned briefly some experiments with Japanese, but gave no results. Differences between languages, and especially differences in resource availability for different languages, can affect the construction of the similarity function at the heart of CBC. CBC also crucially depends on several thresholds whose values were established experimentally. It is quite unclear to what extent they can be reused or re-discovered for different languages and language resources.

The CBC algorithm has been well described by its authors (Pantel, 2003, Lin and Pantel, 2002). We will therefore only outline its general organisation, following (Lin and Pantel, 2002) and emphasising selected key points. We have reformulated some steps in order to name consistently all thresholds present in the algorithm. Otherwise, we keep the original names.

I. Find most similar elements

1. For each word e in the input set E, select k most similar words consid-

ering only e's features above the threshold θ_{MI} of Mutual Information, cf (Manning and Schütze, 2001).

II. Find committees

1. Extract a set of unique word clusters by average link clustering[23], one highest-scoring cluster per list.
2. Sort clusters in descending order and for each cluster calculate a vector representation on the basis of its elements.
3. Going down the list clusters in sorted order, extend an initially empty set C of *committees* with clusters similar to any previously added committee below the threshold θ_1.
4. For each $e \in E$, if the similarity of e to any committee in C is below the threshold θ_2, add e to the set of residues R.
5. If $R \neq \emptyset$, repeat Phase II with C (possibly $\neq \emptyset$) and $E = R$.

III. Assign elements to clusters

For each e in the initial input set E:

1. S = identify $\theta_{T200} = 200$ committees most similar to e,
2. while $S \neq \emptyset$
   (a) find a cluster $c \in S$ most similar to e,
   (b) exit the loop if the similarity of e and c is below the threshold σ,
   (c) if c "is not similar" to any committee in C[24], assign e to c and *remove* from e its features that *overlap* with c's features,
   (d) remove c from S.

CBC has three main phases, marked by Roman numerals in the outline. In the initial Phase I, data that represent the semantic similarity of LUs are prepared. Here, CBC strongly depends on the quality of the applied MSR – the most important CBC parameter – and the MSR is transformed by taking into consideration only some features (the threshold θ_{MI}) and the k most similar LUs.

In the next two phases, the set of possible senses is first extracted by means of committees; next, LUs are assigned to committees. A *committee* is an LU cluster intended to express some sense by means of a cluster vector representation derived from features that describe the LUs included in it. Committees are selected from the

[23] Average link clustering is also referred to as *group-average agglomerative clustering* (Manning and Schütze, 2001).

[24] We interpret this as c's similarity being below an unmentioned threshold θ_{ElCom}.

initial LU clusters generated by processing the lists of the k most similar LUs, see II.1 and II.2. Only the groups dissimilar to other selected groups are added to the set of committees, because the committees should ideally describe all senses of the input LUs, see II.3. The set of committees is also iteratively extended in order to cover senses of all input LUs, see the condition in II.4.

Committees only define senses. They are not the final lemma groups we will extract. The final lemma groups – ideally sets of near-synonyms – are extracted in Phase III on the basis of committees. Each lemma can be assigned to one of several groups by the similarity to the corresponding committees. It is assumed that each sense of a polysemous lemma corresponds to some subset of features which describe the given LU assigned to some committee c (the next sense of e has been identified). CBC attempts to identify the features that describe sense c of e and remove them before the extraction of the other senses of e. The idea behind this operation is to remove sense c from the representation of e, in order to make other senses more prominent.

The original implementation of the *overlap* and *remove* operations is straightforward: the values of all features in the intersection are simply set to 0 (Pantel, 2003). We found this technique too radical. It would be correct if the association of features and senses were strict, but it is very rarely the case. Mostly, one feature derived from lexico-syntactic dependency corresponds in different degree to several senses.

After a manual inspection of data collected in a co-incidence matrix, we concluded that it is hard to expect any group of features to encode some sense unambiguously. Some features also have low, accidental values, while some are very high. Finally, vector similarity is influenced by the whole vector, especially when we analyse the absolute values of similarity by comparing it to a threshold such as. σ in step 2b of the CBC algorithm.

Assuming that a group of features and some part of their *strength* are associated with a sense just recorded, we wanted to look for an estimation of the extent to which feature values should be reduced. The best option seems to be the extraction of some association of features with senses, but for that we need an independent source of knowledge for grouping features, as it was done in (Tomuro et al., 2007). Unfortunately, it is not possible in the case of a language with limited resources like Polish. Instead, we tested two simple heuristics ($w(f_i)$ is the value of the f_i feature, $v_c(f_i)$ – the value of f_i in the committee centroid[25]):

- minimal value:

$$w(f_i) = w(f_i) - \min(w(f_i), v_c(f_i))$$

[25] The centroid features are calculated as average from the features of vectors in the committee.

- the ratio of committee importance:

$$w(f_i) = w(f_i) - w(f_i)\frac{v_c(f_i)}{\sum v_c(\bullet)}$$

In the minimal value heuristic, we make quite a strong assumption that a feature is associated only with one sense on one of the sides: LU or committee. The lower value identifies the right side. The ratio heuristic is based on a weaker assumption: the feature corresponds to the committee description only to some extent.

3.5.4 Benefits of discovered senses for constructing a wordnet

During the reimplementation of CBC for Polish we stumbled upon two problems. There are significant typological differences between Polish and English, and the availability of language tools differs. For example, Polish – unlike English – is generally a free word-order language; much syntactic information is encoded by rich inflection. This makes the construction of even a shallow parser for Polish more difficult than for English – and CBC begins by running a dependency parser on the corpus. As shown in Section 3.4 and in (Piasecki et al., 2007a, Broda et al., 2008), this similar problem can be solved by applying several types of lexico-morphosyntactic constraints. This identifies a subset of structural dependencies mainly from morphosyntactic agreement among words in a sentence and a few positional features like noun-noun head/modifier pairs. The constructed MSR gave results comparable with the results achieved by humans in the same task (Piasecki et al., 2007b). We therefore assumed that the constructed MSR is at least comparable in quality to that used in (Pantel, 2003, Lin and Pantel, 2002). We adopted the constraint-based approach here, applying a subset of lexico-morphosyntactic constraints as in Section 3.4.3: noun modification by a specific adjective or a specific adjectival participle (AdjC), and noun co-ordination with a specific noun (NcC).

Evaluation of the extracted word senses proposed in (Lin and Pantel, 2002, Pantel, 2003) is based on comparing the extracted senses with those defined for the same words in PWN. It is assumed that a correct sense of word w is described by a word group c containing w if a PWN synset s containing w is sufficiently similar to c. The latter condition is represented by another threshold θ.

Similarity between wordnet synsets is central to the evaluation proposed in (Lin and Pantel, 2002, Pantel, 2003). Similarity was defined through probabilities assigned to synsets and derived from a corpus annotated with synsets. This kind of synset similarity is very difficult to estimate for languages for which there is no such corpus, as is the case of Polish. In order to avoid any kind of unsupervised estimation of synset probabilities, we used a slightly modified version of Leacock's similarity measure (Agirre and Edmonds, 2006):

$$sim(s_1, s_2) = -\log\left(\frac{Path(s_1, s_2)}{\max_{s_a, s_b} Path(s_a, s_b)}\right), \tag{3.9}$$

$Path(a, b)$ is the length of a path between two synsets in *plWordNet*.

Except for synset similarity, we follow (Lin and Pantel, 2002, Pantel, 2003) strictly in other aspects of word-sense evaluation. Synset similarity is used to define the similarity between a word w and a synset s. Let $S(w)$ be a set of wordnet synsets including w (its senses). The similarity between s and w is defined as follows:

$$simW(s, w) = \max_{t \in S(w)} sim(s, t) \tag{3.10}$$

The similarity of a synset s (a sense recorded in a wordnet) and a group of LUs c (extracted sense) is defined as the average similarity of LUs belonging to c. LU groups extracted by CBC have no strict limits. Their members are of different similarity to the corresponding committee (sense pattern). The core of the LU group is defined in (Lin and Pantel, 2002, Pantel, 2003) via a threshold κ[26] on the number of LUs belonging to the core. Let also c_κ be the core of c – a subset of κ most similar members of c's committee. The similarity of c and s is defined as follows:

$$simC(s, c) = \frac{\sum_{w \in c_\kappa} simW(s, u)}{\kappa} \tag{3.11}$$

We assume that a group c corresponds to a correct sense of w if

$$\max_{s \in S(w)} simC(s, c) \geq \theta \tag{3.12}$$

The wordnet sense of LU w, corresponding to the sense of w represented by a lemma group c is defined as a synset which maximizes the value in formula 3.12:

$$\arg\max_{s \in S(w)} simC(s, c) \tag{3.13}$$

The question arises why this evaluation procedure is so indirect. Why do we not compare the cores of the LU groups with wordnet synsets? The answer is seemingly simple. Both in Polish and in English, certain matches are hard to obtain. LU groups are indirectly based on the MSR used. They do not have clear limits, and still show some closeness to a sense, but not to a strictly defined sense. On the other hand, wordnet synsets also have a substantial level of subjectivity in their definitions, especially when they are intended to describe *concepts*, which are not directly observable

[26] We changed the original symbol k to κ so as not to confuse it with k in the algorithm.

in language data. The indirect evaluation defined in (Pantel, 2003, Lin and Pantel, 2002) will measure the level of resemblance between the division into senses made by linguists constructing the wordnet and that extracted via clustering.

We wanted to evaluate the algorithm's ability to reconstruct plWordNet synsets. That would confirm the applicability of the algorithm in the semi-automatic construction of wordnets. We put nouns from plWordNet on the input list of nouns (E in the algorithm). Because plWordNet is constructed bottom-up, the list consisted of 13298 most frequent nouns in IPIC plus some most general nouns, see Section 3.4.5. The constraints were parameterised by 96142 features (41599 adjectives and participles, and 54543 nouns).

Several thresholds used in the CBC algorithm (plus a few more in the evaluation) are the major difficulty in its exact re-implementation. No method of optimising CBC in relation to thresholds was proposed in (Pantel, 2003, Lin and Pantel, 2002)[27] and the values of all thresholds in (Pantel, 2003) were established experimentally. There also was no discussion of their dependence on the applied tools, corpus and characteristics of the given language.

Broda et al. (2008) performed such analysis in relation to Polish. Here we will outline only most important conclusions. Experiments with using RWF instead of PMI showed that RWF gives higher precision (38.81% versus 22.37%), but leads to fewer resulting word assigned to groups (744 versus 2980). The value of σ, which controls when to stop assigning words to a committee (step 2b in Phase III of the algorithm), must be carefully selected for each type of MSR separately. As the value of σ increases, the precision increases too, but the number of words clustered drops significantly. When we make σ small and θ_{ElCom} (meaning that word "is not similar" to any committee), we get relatively good precision but more words clustered. We found that contrary to the statement and chart in (Pantel, 2003), tuning both thresholds was important in our case.

The experiments confirmed our intuition that removing overlapping features in Phase III of CBC is too radical. The application of both proposed heuristics was tested experimentally and resulted in the increased precision. The minimal-value heuristic increased the precision from 38.8% to 41.0% on 695 words clustered. The usage of the ratio heuristic improves the result even further: the precision rises to 42.5% on 701 words clustered. A manual inspection of the results showed that the algorithm tends to produce too many overlapping senses when it uses the ratio heuristic.

Because of indirect nature of evaluation proposed in (Pantel, 2003) we wanted to evaluate CBC in more direct and intuitive way. We assumed that proper clustering

[27] Automatising this process is very difficult, because the whole process is computationally very expensive. A full iteration takes 5–7 hours on a 2.13GHz PC with 6GB of RAM, which makes, say, an application of Genetic Algorithms barely possible.

should be able to clear the MSR from accidental or remote associations. That is to say, if two words belong to the same word group, it is a strong evidence of their being near-synonyms or at least being closely related in the hypernymy structure. If that is true then accuracy in WBST+H and EWBST tests (Sec. 3.3) of MSR enriched with output of CBC should be better than MSR alone. That kind of evaluation of CBC was performed by Broda et al. (2008). Accuracy of joined algorithm was lower then for MSR alone. Both methods of evaluation showed that CBC applied for Polish tends to group loosely related lemmas too often. Even improvement in removing overlapping features in Phase III did not yield satisfying precision.

In order to illustrate the work of the algorithm, we selected two examples of correct word senses extracted for two polysemous LUs. The word senses are represented by committees described by numeric identifiers. It is thus emphasised that committee members define only some word sense and are not necessarily near synonyms of the given LU.

LU: **bessa** *economic slump*

id=95 committee: {niezdolność *inability*, paraliż *paralysis*, rozkład *decomposition*, rozpad *decay*, zablokowanie *blockage*, zapaść *collapse*, zastój *stagnation*}

id=153 committee: {tendencja *tendency*, trend *trend*}

LU: **chirurgia** *surgery*

109 committee: {biologia *biology*, fizjologia *physiology*, genetyka *genetics*, medycyna *medicine*}

196 committee: {ambulatorium *outpatient unit*, gabinet *cabinet*, klinika *clinic*, lecznictwo *medical care*, poradnia *clinic*, przychodnia *dispensary*}

Now, the same but with the proposed *heuristic of minimal value activated.*

LU: **bessa**

64 committee: {pobyt *stay*, podróż *travel*} – a spurious sense

95 committee: **as above**

153 committee: **as above**

LU: **chirurgia**

109 committee: **as above**

171 committee: {karanie *punishing*, leczenie *treatment*, prewencja *prevention*, profilaktyka *prophylaxis*, rozpoznawanie *diagnosing*, ujawnianie *revealing*, wykrywanie *discovering*, zapobieganie *preventing*, zwalczanie *fighting*, ściganie *pursuing, prosecuting*} – a correct additional sense found

196 committee: **as above**

Next, two examples of committees and the generated word groups.

- **committee 57**: {ciemność *darkness*, cisza *silence*, milczenie *silence = not speaking*}
- **LU group**: {cisza, milczenie, ciemność, spokój *quiet*, bezruch *immobility*, samotność *solitude*, pustka *emptiness*, mrok *dimness*, cichość *silence (literary)*, zaduma *reverie*, zapomnienie *forgetting*, nuda *ennui*, tajemnica *secret*, otchłań *abyss*, furkot *whirr*, skupienie *concentration*, cyngiel *trigger*, głusza *wilderness*, jasność *brilliance*}
- **committee 69**: {grota *grotto*, góra *mountain*, jaskinia *cave*, lodowiec *glacier*, masyw *massif*, rafa *reef*, skała *rock*, wzgórze *hill*}
- **LU group**: {góra, skała, wzgórze, jaskinia, masyw, pagórek *hillock*, grota, wzniesienie *elevation*, skałka *small rock*, wydma *dune*, górka *small mountain*, płaskowyż *plateau*, podnóże *foothill*, lodowiec, wyspa *island*, wulkan *volcano*, pieczara *cave*, zbocze *slope*, ławica *shoal*}

Finally, an example of a polysemous committee and the lemma group generated on this basis. The group clearly consists of two separate parts: animals and zodiac signs.

- **committee 11**: bestia *beast*, byk *bull*, lew *lion*, tygrys *tiger*
- **LU group**: {lew, byk, tygrys, bestia, wodnik *aquarius*, koziorożec *capricorn*, niedźwiedź *bear*, smok *dragon*, skorpion *scorpio*, nosorożec *rhinoceros*, bliźnię *twin*, lampart *leopard*, bawół *buffalo*}

The last examples clearly show the role of the committee in defining the main semantic axis of the LU group. Two general semantically different LUs in the same committee make it ambiguous between at least two senses. Such a committee results in inconsistent LU groups created from it. Thus the initial selection of committees is crucial for the quality of the whole algorithm, and the quality of CBC depends directly on the MSR applied.

Although the experiments performed with CBC gave interesting and promising results, we achieved too low accuracy for using the output (i.e. the groups of words) as a direct help in expanding plWordNet (not to mention treating the groups as synset proposals). Nevertheless, we plan to improve the CBC accuracy even further in the future, and use it as one additional knowledge source for the semi-automatic expansion of plWordNet.

Chapter 4

Extracting Instances of Semantic Relations

Pattern-based approaches originate from the observation that dictionary definitions in *Machine-Readable Dictionaries* [MRD] follow a limited number of fixed schemes characterised also by a limited range of syntactic constructions. A typical dictionary definition of a word sense describes it by a genus term followed by set of differentiae or a short description referring to related words, e.g. (Matsumoto, 2003). Work on the extraction of lexico-semantic relations from MRD started already in the 1980s (Amsler, 1981), following (Matsumoto, 2003). Dictionary definitions are processed using patterns that identifying selected expressions. Patterns either are regular expressions or or are written in a formal language of a similar expressive power.

In corpora, patterns are used to recognise pairs of LUs as instances of a specified lexico-semantic relation. Consider, for example, a pattern from the seminal work of Hearst (1992):

$$\mathtt{NP}_0 \;\mathtt{...}\;\; \mathtt{such\ as}\; \mathtt{NP}_1,\; \mathtt{NP}_2\; \mathtt{...}\;\; \mathtt{(and\ |\ or\ )}\; \mathtt{NP}_n$$

It implies that each noun phrase $\mathtt{NP}_i$ is a hyponym of the noun phrase $\mathtt{NP}_0$, or, more precisely, the hypernymy relation holds between LUs represented in the text by the given noun phrases. Hearst (1992, 1998) constructed manually only five patterns frequently matched in a corpus and appealingly accurate. The accuracy was measured as a number of LU pairs linked by the hyponymy relation in PWN to all those extracted. For the pattern shown above, for example, 61 of 106 extracted LU pairs from Grolier Encyclopedia were confirmed in PWN (Hearst, 1992).

The implicit assumption here is that one can construct patterns accurate enough to draw correct conclusions from single occurrences of pairs of LUs. In general , however, it seems barely possible due, amongst others, to the presence of metaphor. Without deeper semantic and pragmatic analysis, instances of metaphor may be hard to distinguish from literal uses. Hearst extracted *aeroplane* as a hyponym of *target* and *Washington* as an instance of *nationalist*; such derived associations are clearly specific to particular documents from which they were extracted. Another problem is the scarcity of pattern instances in corpora; merely 46 instances were acquired from 20 million words of the New York Times corpus (Hearst, 1992).

Still, lexico-syntactic patterns are worth attention from the perspective of expanding a wordnet. Their accuracy in identifying instances of wordnet relations is much higher than the accuracy of the MSRlist$_{(x,k)}$ lists (constructed using an MSR) – see the manual evaluation in Section 3.4.5. Each pattern is also focused on instances of one semantic relation, mostly hypernymy. In this way, patterns can be a valuable source of knowledge, complementary to MSRs. This has been the approach in Caraballo (1999, 2001) (Section 4.5.3).

4.1 Lexico-Morphosyntactic Patterns

There are few known applications of the pattern-based paradigm to Polish. Martinek (1997) and Ceglarek and Rutkowski (2006) presented attempts to apply patterns to MRDs. In an application to a corpus, Dernowicz (2007) performed a simple experiment: using lexico-syntactic patterns to extract meronymic and hypernymic pairs for a very limited set of words in a limited domain. The major obstacles are the lack of available electronic dictionaries and problems with preprocessing text in Polish. When processing English, a chunker can pick noun phrases and this reduce the complexity of the text is reduced before patterns are applied. Nearly fixed word order also helps. It is not so in Polish: no chunkers or shallow parsers are available (see the discussion in Section 3.4.3). The word order is much freer; as an example, consider two almost synonymous expressions that both point to the same instance of hyponymy:

wieloryb jest ssakiem (*whale*$_{case=nom}$ *is mammal*$_{case=inst}$)
ssakiem jest wieloryb (*mammal*$_{case=inst}$ *is whale*$_{case=nom}$)

Polish verb arguments are marked by grammatical case, so the morphosyntactic properties of words encode much of the syntactic structure – Section 3.4.3. We can therefore reapply to pattern identification the processing methods presented there. We have built an IPIC subcorpus (we named it HC) with some 80000 sentences that plWordNet signals as containing LU pairs linked by hypernymy. Hearst's patterns (Hearst, 1992) were the first inspiration for discovering patterns for Polish. We also manually examined several hundred sentences from HC, looking for characteristic configurations of LUs. Rather naturally, we were guided by language intuitions. The *Estratto* algorithm (see the next section) automatically extracts patterns and instances; its results were another interesting source of ideas. Although *Estratto* produces rather generic patterns, some such simple patterns became a starting point for the manual development of more refined versions.

A pattern imposes lexico-morphosyntactic constraints on the elements of a text fragment and their mutual dependencies. We implemented the constraints in the JOSKIPI language (Section 3.4.3) , very similarly to how we wrote lexico-morphosyntactic constraints for context description in MSR extraction. Preprocessing by the morphosyntactic tagger TaKIPI was assumed.

In the end, we found five different patterns for extracting hypernymy instances. We only adopted for further stages those patterns that pick out more than a few LU pairs in HC[1]. In the following schematic description, we show (i) the target nominal LUs `NLU1` and `NLU2` with their positions and constraints on the grammatical category values, (ii) the trigger words (`base` is the root, `wf` – the word form), (iii) constraints on LUs that occur in between[2] '. . . ' denotes any sequence of tokens.

***JestInst*:** `NLU1(cas=nom) ...(base=być(`*to be*`)) ...`
`NLU2(cas=inst, nmb=nmb(NLU1))`
— `NLU1` is supposed to be a hyponym and `NLU2` a hypernym; there also is a configuration with the reversed positions of both target LUs and roles signalled by case values (Figure 4.1);

***NomToNom*:** `NLU1(cas=nom) (Adj|Adv|Noun|Num)* (base=to)`
`(`≈ copulative *is*`) (Adj|Adv|Noun|Num)* NLU2(cas=nom)`
— in theory, this spots synonym pairs, but in practice it often pick out `NLU2/NLU1` hypernymy (all other nouns must be in the genitive case);

***IInne*:** `NLU1 (Adj|Adv|Noun|,)* (base∈{i, oraz}(`*and*`))`
`(base∈{inny, pozostały}(`*other*, *remaining*`), nmb=pl)`
`(Adj|Adv)* NLU2(cas=cas(NLU1))`
— similar to Hearst's well-known pattern, this also finds `NLU2/NLU1` hypernymy;

***TakichJak*:** `NLU1(cas≠gen)`
`(Adj|Adv|PartAdj|PartAdv|Noun|Num|Pron|Conj|Punct)*`
`(base=taki (`*such*`)) (base=jak (`*as*`))`
`(Adj|Adv|PartAdj|PartAdv|Noun|Num|Conj|Punct)* NLU2(cas=nom)`
— structurally similar to *IInne*: `NLU2` is one of the hyponyms related to the hypernym `NLU1`; for details of the part between `NLU1` and *taki* see Figure 4.2;

***WTym*:** `NLU1(cas≠gen) (Adj|Adv|Num){0,2} (base=,) (base=w) (wf=tym)`
`(Adj|Adv|Num){0,2} NLU2(cas=cas(NLU1))`
— another version of *IInne*:`NLU1` is a hypernym and `NLU2` is one of the hyponyms.

In view of the the overall goal of semi-automatic expansion of plWordNet, we did not apply patterns to corpus completely freely. Instead, we have specified the

[1] Some rejected patterns may be more prolific in a larger corpus, but we focused on the sensitivity to plWordNet's understanding of hypernymy.

[2] Pattern names contain the Polish trigger words.

target nominal LUs not only with some constraints by also with the pairs of LUs in focus. That is why we used here the same set of 13285 nominal lemmas which were selected for the construction of nominal MSR and the expansion of the core plWordNet (Section 3.4.5). From the set, we generated all possible pairs. We also reapplied the mechanism of co-incidence matrix construction. Target LUs were assigned to rows, and patterns with the position `NLU2` instantiated to subsequent target LUs were assigned to columns. The patterns were run with position `0` representing `NLU1`[3].

Given these assumptions, there is no need to test the presence of `NLU1` in the *IInne* and *TakichJak* code (Figures 4.1–4.2). We refer the reader to Section 3.4.3 for the details of JOSKIPI. *IInne* is implemented in two symmetrical parts joined by `or` for two configurations of the hyponym (`NLU1`) and hypernym (`NLU2`). The matrix construction requires that we start with the hyponym in position `0`. In the first part, we first test if the potential `NLU1` is nominative, then look to the right (till the end of the sentence) for the first verb word form or the first nominal LU and record its position in variable `$X`. We test if it is a form of the verb `być` (*to be*) – any other verb or noun means that the sentence does not match the pattern. We look further to the right for the first verb or the first nominal LU, or a preposition (`prep`) that requires the instrumental case. The latter is necessary, because `NLU2` in the pattern is only identified by the case value induced by the verb `być`. The token at position `$Y` is compared with the base form with which the pattern was instantiated[4]. We also test its case and number.

In the pattern *TakichJak* in Figure 4.2, the iteration goes in the opposite direction. Hyponyms now follow the hypernym, and we wanted to keep the same ⟨hyponym, hypernym⟩ order of the extracted LU pairs across all the patterns. After the case of `NLU1` has been tested, we look to the left till the beginning of the sentence for the sequence `taki jak` (*such as*). Next, we test the tokens between `0` and `$+2T` – the position after `jak` – for the presence of `only` LUs of the specified grammatical classes plus the specified punctuation marks and conjunctions; this signal a coordinate sequence of noun phrases. Finally, `NLU2` is sought further to the left, and tokens between it and `taki` are tested. Only modifiers are accepted there, including nouns and pronouns in the genitive case.

The implementation of the other three patterns is similar.

The patterns *IInne*, *WTym* and *TakichJak* are structurally very similar: a hypernym and a list of hyponyms. Also, a preliminary evaluation on a part of IPIC showed

[3] Multiword LUs were recognised during preprocessing and folded into a one-token representation with the attribute and root set to the values proper for the whole LU. During matrix construction, each target LU occupies exactly one token in the preprocessed representation of the corpus (Broda and Piasecki, 2008b). Recognition of multiword LUs was limited to target LUs (all parts of speech) due to the labour-intensity of their syntactic description.

[4] Technically, each column in the matrix is assigned its own copy of the pattern instantiated to the appropriate nominal LU as `NLU2`.

```
or(
 and(
  equal(cas[0],{nom}),
  rlook(1,end,$X, inter(flex[$X],{adjectival participles, noun,
                                   pronouns, verbal grammatical classes}
  ) ),
  equal(base[$X],{"być"}),
  rlook($+1X,end,$Y,or(
                  inter(flex[$Y], {adjectival passive participle,
                                   noun, pronouns, verbal grammatical classes}),
                  and( equal(flex[$Y],{prep}),
                       equal(cas[$Y],{inst})
                  )
                 )),
  inter(flex[$Y],{subst,ger,depr}),
  equal(base[$Y],{"NP2"}),
  equal(cas[$Y],{inst}),
  equal(nmb[$Y],nmb[0])
 ),
 a symmetrical condition for the right context
)
```

Figure 4.1: The essentials of the *JestInst* pattern implementation in JOSKIPI

that they have very similar accuracy. That is why we decided to merge them into a complex pattern that combines the constraints using the `or` operator. We will refer to this pattern as *mIInne* – see, for example, Table 4.1.

4.2 Benefits of Handwritten Patterns for Wordnet Expansion

We ran experiments on the extraction of hypernymic pairs on the same three corpora as those used for MSR extraction (Section 3.4.5): the IPI PAN Corpus [IPIC] (≈ 254 million tokens) (Przepiórkowski, 2004), the *Rzeczpospolita* corpus [RzCorp] (≈ 113 million tokens) (Rzeczpospolita, 2008), and a corpus of large texts in Polish from Internet (≈ 214 million tokens) [WebCorp]. Table 4.1 presents detailed results for three patterns, *JestInst*, *NomToNom* and *mIInne*.

We assessed the accuracy manually on randomly selected samples. Similarly to other manual evaluations (for example, Section 3.4.5), we determined sample sizes following the method discussed in (Israel, 1992), aiming for the 95% confidence level on the whole population. We used a program named *Sprawdzacz* (Kurc, 2008) that facilitates manual evaluation of the extracted lexico-semantic relation instances[5].

[5] We thank Roman Kurc for his great help with the whole plWordNet project.

```
and(
  in(cas[0],nom),
  llook(-1,begin,$T,equal(base[$T],{"taki"})),
  equal(base[$+1T],{"jak"}),
  only($+2T,-1,$AR,or(
                    inter(flex[$AR],{adjective , adjectival participles , adverb ,
                                     adverbial participles , noun ,numeral }),
                    in(orth[$AR],{"i","lub","czy","oraz","a",",",":","(",")"})
                   )),
  llook($-1T,begin,$N,and(
                        inter(flex[$N],{noun}),
                        equal(base[$N],{"base form of NLU2"}),
                        in(cas[$N],{nom,acc,dat,inst,loc,voc})
                      )),
  only($+1N,$-1T,$AL,or(
                      inter(flex[$AL],{adjective , adjectival participles , adverb ,
                                       adverbial participles , numeral}),
                      and(
                        inter(flex[$AL],{noun , pronouns}),
                        equal(cas[$AL],{gen})
                      )
                     ))
)
```

Figure 4.2: The essentials of the *TakichJak* pattern implementation in JOSKIPI

During the evaluation, an extracted LU pair could be classified as a correct instance of hypernymy (possibly indirect, with longer paths accepted), or as one of two forms of nearly correct instances:

- not the expected hyponym/hypernym order; such pairs occurred more often among the results of the *NomToNom* pattern in which the direction is not marked by grammatical case;

- small inaccuracies in one of the LUs: it is part of a larger multiword LU, or it has a wrong number value, or it is represented by a wrong root (a tagger error).

All other pairs were classified as incorrect. The results in Table 4.1 have been calculated with the assumption that correct and nearly correct instances are positive. If we excluded the nearly correct class, the results would be about 20% lower. The results would be very low if we only sought direct hypernymy. This clearly suggests that the extracted pairs are not directly helpful in expanding the core plWordNet, but they still are a valuable source of knowledge. They show not only semantic similarity of the LUs in a pair, but also the direction of the relation. Indirect hypernyms can be helpful in identifying the right place for a new LU in the existing hypernymy structure. This is how we will use them in the wordnet expansion algorithm presented in Section 4.5.3.

	IPIC		WebCorp		RzCorp		*all*	
Pattern	No.	Acc.	No	Acc.	No	Acc.	No	Acc.
JestInst	60880	11.61	44888	11.97	30063	12.42	121730	10.89
NomToNom	10404	13.5	6414	15.43	4465	14.85	20310	15.66
mIInne	14611	30.06	5983	32.52	6682	33.16	24437	30.69
in 2 patt.	—		—		—		8777	41.05
in 3 patt.	—		—		—		620	74.03

Table 4.1: The results of hypernymy instance extraction by manually constructed lexico-morphosyntactic patterns (No. is the number of LU pairs extracted, Acc. – the accuracy [%], in i patt. – LU pairs occurring in the results of at least i patterns)

The accuracy increased when we applied the patterns only to the closed list of target nominal LUs. There was lower accuracy of the acquired LU pairs in the preliminary experiments on the practically unlimited set of the target nominal LUs, acquired from the joint corpus.

The last two rows in Table 4.1 present the results of voting based on the three patterns applied to the joint corpus. The accuracy doubles in relation to *mIInne*, which produces the best result among all three patterns used alone (in the case of voting when we request the confirmation of an LU pair by all three patterns). On the other hand, the number of LU pairs covered by two or three pattern drops sharply in relation to the list produced by the subsequent patterns. However the number of LU pairs covered by two or three patterns drops sharply in relation to the list produced by the subsequent patterns. The voting experiments showed that the lower-accuracy patterns *JestInst* and *NomToNom* can help increase the final accuracy when combined with *mIInne*.

The corpora used seem to be independent of the number of unique LU pairs extracted by all patterns. In all three cases the number of pairs extracted from the joint corpus is almost the sum of the numbers for the contributing corpora[6]. Still, not every corpus appeared to be an equally good basis for the application of patterns.

It is hard to find a correlation between the frequencies of the extracted LU pairs and their accuracy, especially for *JestInst*. High frequencies (> 100) are produced by collocations, and a typical frequency of a pair is 1–3. They are too low for statistical evaluation. A potential evaluation should take into account the statistical properties of the LUs and the pairs. Such a mechanism has been proposed in the literature for extraction based on automatically generated generic patterns. We will discuss it in Section 4.3.

[6] We tried to make the corpora free of duplicated texts (some duplication seems unavoidable), and there were – to some extent – various genres, but the results were still surprising.

Metaphor is a major source of errors and, even more so, are relations between larger noun phrases, which the patterns assign only to the heads. A typical situation: *NomToNom* captures `NLU2` that includes a relative clause, but only the head is considered. Even a nominal modifier in genitive or an adjectival modifier often makes the meaning of the noun phrase different from the lexical meaning of the head. The conditions in *mlInne* do not constrain the case of the nominal LUs, so it is quite common to erroneously recognize hyponymy for a noun in genitive that is not the head. It is not easy, however, to identify complex Polish noun phrases in genitive. The error rate would be cut if we could apply a good chunker or even a shallow parser combined with the analysis of the meaning relations between structurally related noun phrases – see, for example, (Jacquemin, 2001).

Examples of LU pairs extracted by all three patterns appear in Figure 4.3. Figure 4.4 presents examples of LU pairs extracted from the joint corpus by each of the three patterns.

The results of the application of lexico-morphosyntactic patterns are valuable, but there remains an impression that more could be achieved by following the main line of the pattern-based paradigm. We will now shift our attention to approaches to automatic extraction and evaluation of more generic patterns.

4.3 Generic Patterns Verified Statistically

A manual construction of lexico-syntactic patterns is not laborious if we rely more on intuition than on an intensive survey of known hypernymy instances and the context of their occurrences in a corpus. Morin and Jacquemin (1999) proposed semi-automated discovery of lexico-syntactic patterns. Given a predefined list of hypernymy instances, sentences including these LU pairs are extracted and transformed into "lexico-syntactic expressions". Next, common environments that generalise the expression are produced by considering the similarity of the expressions and a generalisation procedure: lexico-syntactic patterns describing commonalities of expression subgroups are deduced. The pattern extraction procedure still assumes manual verification of the deduced patterns, and the patterns are next applied without automatic evaluation of their accuracy and the reliability of the extracted pairs. The latter is especially important for the application of generic (weakly constraining) patterns to large corpora.

Manually constructed patterns are claimed to have good precision but very low recall (Hearst, 1998). Recall can be increased by using more generic patterns extracted automatically from a corpus, with broad coverage but intrinsically low precision.

Most of the proposed methods follow the common scheme: given the initial examples of the target relations, henceforth called *seeds*, patterns are generated from the corpus and next used to extract further instances. Methods differ in pattern generation

Correct hypernymy instances	
koncesja (*concession*)	decyzja (*decision*)
kapłan (*priest*)	człowiek (*human*)
maj (*May*)	okres (*period*)
kwestia (*issue*)	problem (*problem*)
sowa (*owl*)	ptak (*bird*)
klient (*customer*)	osoba (*person*)
pielęgniarka (*nurse*)	osoba (*person*)
profesor (*profesor*)	człowiek (*human*)
galeria (*gallery*)	miejsce (*place*)
matematyka (*mathematics*)	przedmiot (*subject*)
matka (*mother*)	kobieta (*woman*)
helikopter (*helicopter*)	maszyna (*machine*)
droga (*way*)	szlak (*track*)
zespół (*team*)	grupa (*group*)
mecz (*game*)	spotkanie (*meeting*)
restrukturyzacja (*restructurisation*)	zmiana (*change*)
konsument (*consumer*)	osoba (*person*)
tenis (*tennis*)	sport (*sport*)
festiwal (*festival*)	impreza (*event*)
dziennik (*daily*)	dokument (*document*)
medycyna (*medicine*)	nauka (*science*)
anioł (*angle*)	istota (*being*)
spółka (*partnership*)	firma (*firm*)
szczur (*rat*)	szkodnik (*pest*)
skorpion (*scorpio*)	znak (*sign*)
rak (*cancer*)	choroba (*illness*)
nagroda (*prize*)	wyróżnienie (*distinction*)
Non-hypernymy associations	
przepis (*recipe*)	kwestia (*issue*)
silnik (*engine*)	jednostka (*unit*)
człowiek (*human*)	drzewo (*tree*)
program (*program*)	działanie (*activity*)
muzyka (*music*)	dźwięk (*sound*)
istota (*being*)	nic (*nothing*)
wojsko (*army*)	organizacja (*organisation*)
stowarzyszenie (*association*)	instytucja (*institution*)
cień (*shadow*)	wróg (*enemy*)
książka (*book*)	materiał (*material*)
słońce (*sun*)	czynnik (*factor*)

Figure 4.3: Examples of LU pairs extracted from the joint corpus by all three lexico-morphosyntactic patterns

JestInst		Corr.
zawodnik (*contestant*)	osoba (*person*)	C
pałka (*baton*)	broń (*weapon*)	C
powód (*plaintiff*)	osoba (*person*)	C
antologia (*anthology*)	dzieło (*work*)	C
pszczoła (*bee*)	zwierzę (*animal*)	C
Ormianin (*Armenian*)	mieszkaniec (*inhabitant*)	C
koncert (*concert*)	utwór (*composition*)	C
człowiek (*human*)	kurier (*courier*)	C
brat (*brother*)	przeciwieństwo (*opposition*)	I
ludobójstwo (*genocide*)	słowo (*word*)	I
artystka (*artist*$_{fem.}$)	uczennica (*pupil*$_{fem.}$)	I
spółka (*partnership*)	członek (*member*)	I
NomToNom		
dochód (*income*)	przychód (*revenue*)	C
dowódca (*commander*)	człowiek (*human*)	C
ocena (*evaluation*)	informacja (*information*)	C
stomatolog (*stomatologist*)	lekarz (*physician*)	C
kac (*hangover*)	objaw (*symptom*)	C
premier (*premier*)	polityk (*politician*)	C
sesja (*session*)	część (*part*)	C
dokument (*document*)	notatka (*memo*)	C
mąż (*husband*)	pijak (*drunkard*)	I
cierpienie (*suffering*)	abstrakcja (*abstraction*)	I
rozmowa (*conversation*)	strata (*loss*)	I
socjalizm (*socialism*)	pustka (*emptiness*)	I
mIInne		
regulacja (*regulation, adjustment*)	zasada (*principle*)	C
krzyż (*cross*)	znak (*sign*)	C
jaskinia (*cave*)	miejsce (*place*)	C
podawanie (*serving*$_{gerund}$)	czynność (*activity*)	C
poker (*poker*)	gra (*game*)	C
spółka (*partnership*)	podmiot (*subject*)	C
naród (*nation*)	lud (*people*)	C
chrześcijaństwo (*Christianity*)	religia (*religion*)	C
obliczenie (*calculation*)	dokument (*document*)	I
kredyt (*credit*)	koszt (*cost*)	I
wyposażenie (*equipment*)	czynnik (*factor*)	I
przebieg (*course (of)*)	szczegół (*detail*)	I

Figure 4.4: Examples of LU pairs extracted by different lexico-morphosyntactic patterns from the joint corpus (C – correct, I – incorrect)

and in the evaluation of the extracted patterns and instances (the evaluation is not present in some methods). Very often the process is recursive: the extracted instances are used to generate a new pattern set.

Brin (1999) proposed a method of extracting patterns that discover *author-book* pairs in Web pages. Pair occurrences are represented by the order of both names and by *prefix*, *middle* and *suffix* character context. Patterns are generated by first grouping seed occurrences by the order and the middle. For each group, the longest matching prefix and suffix are identified, and one pattern for a group is extracted. Patterns are evaluated by their specificity defined as the product of the lengths of the prefix, middle and suffix. Pattern with specificity below a predefined threshold are rejected[7]. Brin did not present any thorough evaluation or any accuracy data.

Agichtein and Gravano (2000) as well as Agichtein et al. (2001) follow Brin's approach, extended with respect to the recognition of an unlimited set of relations between *Named Entities*, and to the generation and evaluation of patterns and extracted instances. Their system has been aptly named *Snowball* to reflect the iterative character of the algorithm. It starts with the seeds and an empty set of extraction patterns. In each round, it extracts new patterns and new set of instances but keeps only those that have been evaluated as sufficiently reliable. The previous set of instances becomes the seed set for the next iteration.

The text is first processed by a named-entity tagger[8]. During pattern generation, text fragments including pairs of named entities in focus are extracted. For each unique pair, the left, middle and right contexts are represented as vectors of weight in $\langle 0, 1 \rangle$. The weights, produced from term frequencies, are meant to express the importance of the term for the context. Weights for the middle context are higher to reflects its larger importance for the relation representation. The size of the left and right context is fixed to a specified text window.

During extraction, a text fragment around two named entities is transformed to a vector and compared with pattern vectors. It is accepted as an instance if the similarity of vectors (measured as a product) exceeds the threshold. Extracted patterns and instances are evaluated by confidence. The *confidence* of a patterns is the ratio of positive to negative matches, first measured for the initial instances. Pattern confidence for the next iteration is a combination of the new and old value. The confidence of an instance directly depends on the confidence of the patterns that select it and on the degree of matching between the instance and particular patterns. After each iteration, all instances with low confidence (below a threshold) are discarded.

Ravichandran and Hovy (2002) developed a weakly supervised algorithm for extraction of question-answer pairs of named entities, based only on seeds. They used

[7] We have intentionally omitted the role of the URL addresses in pattern generation; we focus on a plain-text corpus as a source.

[8] The MITRE Corporation's Alembic Workbench (Day et al., 1997) was used in Snowball.

a simple tokeniser and a simple sentence boundary recogniser, rather than advanced tools like the named-entity tagger in Snowball. This relaxation of assumptions made their algorithm more general. Patterns are extracted from sentences including seed occurrences by means of suffix trees for extracting substrings of optimal length. Pattern precision is calculated as the ratio of the correctly matched instance occurrences to all matches of the pattern. Instances are ordered by the precision of the patterns selecting them. The process is not iterative.

Pantel et al. (2004) proposed an algorithm for mining *is-a* relations from huge text corpora. Text is first processed by a part-of-speech tagger (Brill, 1995) and stored in a two-level format: surface word forms and part-of-speech tags. Next, all sentences including seeds are extracted. Patterns are learned from the sentences by calculating the minimal edit distance among sentences and registering the edit operations required. Patterns with relatively high occurrence and high precision are identified using the log-likelihood principle (Dunning, 1993) for scoring. Only the 15 highest-score patterns have been used to extract hypernymy instances.

Pantel and Pennacchiotti (2006) proposed a system called *Espresso* which seems to combine all interesting properties of its predecessors. It does not make any assumptions concerning the relation described by the patterns. It works on plain text, uses only a part-of-speech tagger and a simple chunker, and works iteratively during the subsequent phases of pattern and instance extraction and evaluation. It is also claimed to be weakly supervised, requiring only the initial set of seeds. Taking into account the foregoing selective overview of the previous algorithms and the results of the evaluation of Espresso (Pantel and Pennacchiotti, 2006), we decided to use Espresso as the starting point for the development of an algorithm that supports the expansion of the core plWordNet.

Espresso

Espresso follows the *bootstrapping* paradigm in a version exemplified already in the Snowball system (Agichtein and Gravano, 2000). Seeds are used to extract the first set of patterns; the subsequent phases of instance and pattern extraction go on automatically. The following four main phases can be identified in Espresso.

1. *Preprocessing*: the input text is divided into tokens (some multiword expressions are identified) and run through a part-of-speech tagger.

2. *Pattern induction*: sentences including seeds are extracted and patterns are learned using the algorithm in (Ravichandran and Hovy, 2002).

3. *Pattern selection*: extracted patterns are statistically evaluated and ranked by instances inducing them; k top patterns are selected.

4. *Instance extraction*: the selected patterns are used to extract instances; the instances are next statistically evaluated using the patterns that match their occurrences; the m top instances are selected and kept for the next iteration – a possible *expansion phase* for extended retrieval of instances can take place before the selection.

The first step is performed once at the beginning. Lacking a stop condition (it depends on the number of extracted patterns and average pattern score decrease in relation to the previous iteration), the next iteration starts from step 2.

Preprocessing used the Alembic Workbench part-of-speech tagger (Day et al., 1997) but no shallow parser. Multiword terms (if left unrecognised) would decrease the accuracy of the extraction algorithm, because of instances generated from parts of the complex terms. We noticed this problem in the case of manually constructed patterns discussed in Section 4.1. Instead of using a shallow parser, Espresso applied a definition of multiword terms as a regular expression (Pantel and Pennacchiotti, 2006, p. 115). This simple solution cannot be directly ported to languages typologically different from English, such as the morphology-rich, flexible word-order Polish. Morphology and word-order flexibility will be discussed shortly in the context of a proposed modification of Espresso named *Estratto*.

Pattern induction is based on the algorithm of Ravichandran and Hovy (2002), discussed earlier, with only one modification: all recognised multiword terms are replaced with the label *TR*.

The statistical *reliability measures*, introduced in Espresso for ranking patterns and instances, follow the same basic scheme of recursive dependency of both measures: the reliability of instances depends on the patterns which extracted them and the other way around. This scheme can be traced back to Snowball (Agichtein and Gravano, 2000) and (Ravichandran and Hovy, 2002), but there it was implemented in a less sophisticated form. Statistical evaluation is clearly this element which is missing when working with manually extracted patterns. Reliability calculation is Espresso's key element, so we present it in more detail. For the needs of ranking and selecting patterns, Espresso introduced a reliability measure:

$$r_\pi = \frac{\sum_{i \in I}(\frac{pmi(i,p)}{\max_{pmi}} * r_t(i))}{|I|} \tag{4.1}$$

p is a pattern, i – an instance, r_t – a reliability measure for instances, pmi – the Pointwise Mutual Information [PMI] measure, explained below, and $|I|$ – the size of the instance set.

The reliability of each seed delivered to Espresso is 1, so pattern reliability is proportional to the average strength of association between the pattern and the seeds measured by PMI. Later, associating a pattern with a larger number of more reliable instances increases the pattern's reliability.

The reliability of instances is defined symmetrically: replace the reliability of instances is with the reliability of patterns $r_\pi(i)$ and the set of instances by the set of patterns P:

$$r_\pi = \frac{\sum_{p \in P} \left(\frac{pmi(i,p)}{\max_{pmi}} * r_\pi(p) \right)}{|P|} \tag{4.2}$$

PMI originates from Information Theory. It measures the strength of association between two events:

$$pmi(i,p) = \log \frac{|x,p,y||*,*,*|}{|x,*,y||*,p,*|} \tag{4.3}$$

$|x,p,y|$ is the number of occurrences of x and y in contexts matching the pattern p, $x,*,y$ – the number of co-occurrences of x and y in the corpus regardless of the pattern, and so on.

The definition of pmi presented by Pantel and Pennacchiotti (2006) does not include the constituent $|*,*,*|$ (the number of contexts). The PMI measure, however, should be usually greater than 0, while pmi defined in (Pantel and Pennacchiotti, 2006) is not. The missing constituent is also suggested by the general definition of PMI:

$$pmi(i,p) = \log \frac{p(I,P)}{p(I)p(P)} \tag{4.4}$$

Because PMI is significantly higher when instances and patterns are not numerous (e.g., < 10), PMI is multiplied by a discounting factor proposed in (Pantel and Ravichandran, 2004) that decreases the bias towards infrequent events.

In Espresso, generic patterns are defined as generating 10 times more instances than previously accepted reliable patterns. They extract many instances but are characterised by lower reliability. Generic patterns are not excluded by definition. They increase recall (the number of correct instances extracted), but inevitably decrease precision. In order to prevent an excessive reduction of the precision, an additional measure of confidence of instances has been introduced. It is based on the evaluation of instances against reliable patterns only and the additional data acquired by searching the Web with the queries generated from instances and patterns:

$$S(i) = \sum_{p \in P_R} S_p(i) * \frac{r_\pi(p)}{T} \tag{4.5}$$

P_R is the set of reliable patterns (given a threshold), $S_p(i,p)$ is the PMI between i and p measured on the data acquired from the Web (using Google queries) and T is the sum over the reliability of reliable patterns.

Only instances with the confidence measure above a threshold are selected for the next iteration and used to induce and evaluate patterns. The implicit assumption is that the confidence measure can be calculated for the majority of instances: reliable patterns, which are less frequent but more specific, should occur at least a few times with the majority of correct instances on the Web. We discard all correct instances which are not covered by the Web data matching reliable patterns.

It is worth emphasizing that the evaluation of patterns in the next iteration is based on instances used to induce these patterns – not on instances which they extracted.

The intuition behind the measures of reliability and confidence is that patterns which describe the given relation well often occur with many confident instances of this relation – and the other way around. The difference is that instances extracted by generic patterns will get high confidence if they occur in contexts matched by the specific patterns of good reliability in the validating corpus.

The measures of reliability and confidence reduce the need for manual supervision once Espresso has started. In a way they define the degree to which the extracted instances express the target relation and the patterns describe contexts. The former is an advantage over manual patterns, for which collected frequencies are mostly low and accidental, and say a little about the quality of instances.

The system of measures, instances and pattern selection are universal and do not refer to any properties of any particular relation being extracted. Espresso can therefore be applied to a wide range of relations. It was applied to hypernymy, meronymy and antonymy, as well to more specific relations such as person-company or person-job title (Pantel and Pennacchiotti, 2006).

To sum up, Pantel and Pennacchiotti (2006) list Espresso's characteristics:

- high recall together with a small decrease in the precision of extracted instances,
- autonomy of work (weakly supervised algorithm) – only several instances of the given relation must be defined at the beginning,
- independence of the size of the corpus or domain used,
- a wide range of relation types that can be extracted.

Estratto

Estratto (Kurc, 2008, Kurc and Piasecki, 2008) is a modification of Espresso developed mainly to cope with the significant differences between English and Polish: rich morphology, flexible word-order and the much more limited size and access to the Web resources. Let us first present one language-independent adjustment.

Two issues are unclear in how the reliability and confidence measures work. First, reliability is sensitive to fluctuations in PMI values. Higher values (e.g., the effect of small frequencies, even after discounting dependent on the number of occurrences) can cause lower assessment of patterns with a balanced ratio of co-occurrence with matched instances versus the pattern occurrences and instance occurrences alone. Such situations result in artificially increased values of $\max_{pmi}$. We would like to look for a measure which would be less sensitive to the low frequency of pattern matches or instances matched. Also, the value 1 of pattern reliability is not guaranteed even for a pattern which occurs only with a subset of seeds, because of the $\max_{pmi}$ value which can be increased by some infrequent pattern. That is why the propagation of reliability to the subsequent iterations causes new values (calculated for patterns from instances and *vice versa*) to become gradually lower for the respective set. We seek a measure of reliability which returns 1 as the value for the best patterns or instances in every iteration.

$$r_\pi(p) = \frac{\sum_{i \in I} (pmi(i,p) * r_t(i)) * d(I,p)}{\max_p \left(\sum_{i \in I} (pmi(i,p) * r_t(i))\right) * |I|} \tag{4.6}$$

$d(i,p)$ defines how many unique instances the given pattern is associated with.

PMI in formula (4.6) is usually also modified by a discounting factor.

The proposed modifications are intended to increase the reliability of the patterns, which not only extract a lot of instances, but occur with a large number of different instances. The modified measure, when applied to instances, promote those which occur often in the corpus associated with many different patterns.

The choice of the pattern structure is crucial for their expressiveness and the ability to capture those elements of the language structures that express semantic relations between LUs (such as the linear order of constituents in English), but case-marking of noun phrases in Polish (their linear order is mostly insignificant for the potential lexical semantic relation between them). Espresso follows roughly the scheme proposed by Hearst (1992): patterns are regular expressions, in which the alphabet includes word forms and the label *TR* for any multiword term, and a set of variables for noun phrases matched as elements of an instance. The role of part-of-speech tags is unclear in the approach of Pantel and Pennacchiotti (2006), but they are present in the example of the generalisation of a sentence [p. 115]:

Because/IN *TR is*/VBZ *a*/DT *TR and*/CC **x** *is*/VBZ *a*/DT **y**.

We assumed that patterns are simplified regular expressions, with the Kleene closure but without grouping. The alphabet for an inflectional language like Polish should rather include roots than (numerous) word forms. Espresso patterns rely to some extent on the positional, linear syntactic structure of an English sentence. Porting to a significantly different language may be problematic.

The unmarked order in a Polish sentence is Subject-Verb-Object, so simple lexico-syntactic patterns might work similarly as in English. Anything more complicated, such as rich morphosyntactic agreement or even slightly relaxed word/phrase order (usually meaning-preserving) need additional work. We put more emphasis on the morphosyntactic description of pattern elements in terms of the tagset in the IPI PAN Corpus [IPIC] (Przepiórkowski, 2004). The categories include a finer-grained list of parts of speech and dozens of values of several grammatical categories (case, number, gender, person, degree, tense, aspect).

Multiword LU occurrences also get morphosyntactic descriptions – see Section 3.4.5 how this worked with MSR extraction. The linear order of LUs in a Polish sentence need not be correlated with their role in asymmetric lexico-semantic relations. For example, many patterns mark the hypernym and the hyponym by different cases, while their relative positions change. We could generate specific patterns for all different combinations, but we can also look for generalization of a group of patterns based on the morphosyntactic properties.

Following Espresso, patterns are flat and describe a sentence as a sequence of word forms or at most groups of word forms. They are not based on any deeper description of the syntactic structure. The alphabet comprises three types of symbols:

- an asterisk symbol * indicating place for zero or more tokens,
- *root*,
- and markers of *matching locations*, variables in Espresso, which require more structure in Estratto.

The empty symbol represents any LU (represented by any of its word forms). The root represents an actual basic morphological word form of an LU and its grammatical class. This takes care of the likely ambiguity. A matching location includes a partial morphosyntactic description (a reduced version of the IPIC morphosyntactic tag, with selected category values) that represents all LUs with a matching morphosyntactic description. Grammatical classes in IPIC are too fine-grained. We introduced “macro” symbols, such as `noun` that represents jointly all grammatical classes: *substantives*, *gerunds*, *foreign nominals* and *depreciative nouns*.

As in Espresso, there are always two matching locations, at the beginning and the end of a pattern. Patterns do not describe the left and right context of a potential instance.

Matching locations also encode the roles of both LUs identified, for example:

```
(hypo:subst:nom)   jest (hyper:subst:inst)
(hyper:subst:inst) jest (hypo:subst:nom)
```

`jest` = $be_{number=sg,person=3rd}$, `hypo` = hyponym, `hiper` = hypernym, `subst` = *substantive*, `nom` and `inst` are case values, nominative and instrumental.

Pantel and Pennacchiotti (2006) write that patterns can be induced by any pattern-learning algorithm, but only the longest common substring algorithm proposed by Ravichandran and Hovy (2002) was used. The same algorithm was the basis for the generalisation and unification of patterns in Estratto. The algorithm is heuristically guided by a predefined list of relation-specific LUs. Hypernymy, for example, can be signalled by `być` (*be*), `stać się` (*become*), `taki` (*such*), `inny` (*other*), and so on.

In Espresso, the inferred patterns are then generalized by replacing all multiword terms (subsets of noun phrases) by the *TR* labels. Such for Polish might be unworkable: a robust definition of a multiword term as a regular expression seems unattainable – not to mention the lack of a chunker for Polish. As a slightly different method, matching locations are specified via morphological similarity to contexts (partial morphological specification: part of speech and values for the selected grammatical categories), and via predefined relation-specific LUs.

The instance extraction phase follows patterns induction and selection. An instance is a pair $\langle x, y \rangle$ of LUs – instances of the target semantic relation. The authors of Espresso suggest that, given a small corpus, two methods can be used to enrich the instance set. First, each multiword LU in an instance can be progressively simplified down to the head, for example, *new record of a criminal conviction* → *new record* → *record*. A new instance is created the simplified first LU and the second LU intact. Second, a pattern is instantiated only with either x or y, and new instances are retrieved from the Web or an additional large corpus. For example, given the pair (*dog*, *animal*) and the Estratto pattern

```
(hypo:subst:nom) is a/an (hyper:subst:inst)
```

we create two queries:

```
dog              is a/an (hyper:subst:inst)
(hypo:subst:nom) is a/an animal
```

Instances collected using both these methods are added to the instance set. Let us note that in all experiments described by Pantel and Pennacchiotti (2006) only one-word LUs have been used, and the corpora were presumably large enough to provide statistical evidence.

Generalized patterns we described are not considered *generic* so long as they do not generate ten times more instances than the average number of instances extracted by reliable patterns from the previous iteration. High recall, however, results in decreased precision, so every instance extracted by a generic pattern is verified. The verification process starts with instantiating all non-generic patterns with the instances to be verified. The resulting patterns are run on a validating large corpus (for Espresso, the Web). A confidence measure is computed from the collected frequencies and compared with a threshold.

In Espresso, the Internet served as a validating corpus for instances extracted by generic patterns. We need other resources because of the paucity of Polish Web pages and the inherent difficulty of querying regardless of the inflectional variety. A second large corpus (Rzeczpospolita, 2008) (but still much smaller than the data from the Web and even IPIC used for the extraction of the patterns) served the purpose of validation in Estratto. The necessary condition for finding occurrences of the patterns extracted from the primary corpus seems to be that the validating corpus cover similar genres and domains.

The induction of patterns and the extraction of instances in Estratto are controlled by the following set of parameters:

1. the *number of top k* patterns not to be discarded (preserved for the next iterations),
2. the *threshold* measure of confidence for instances,
3. the *minimum frequency* and *maximum frequency* values for patterns,
4. the *minimum size* of a pattern – all patterns that consist of only matching locations and conjunctions are discarded by definition,
5. a *filter* on common words in instances and on instances with identical LUs on both positions,
6. the *size* of the validating corpus.

4.4 Benefits of Extracted Patterns for Wordnet Expansion

We investigated the use of algorithms like Espresso in order to find method for extracting valuable instances of wordnet relations, at least hypernymy, with precision higher than afforded by the handwritten lexico-morphosyntactic patterns. We did not expect ready-to-add hypernymy instances. We only wanted to construct yet another source of knowledge that suggests hypernymy occurrences and the correct direction of the relation.

It is far from trivial properly to evaluate the extracted lexico-semantic resources (Section 3.3). It is much easier for lists of instances: we verify how many of them are correct. Only two comparisons were possible for Polish:

- with the existing structure of the core plWordNet,
- with an evaluation by one of the co-authors.

The former introduces a bias – plWordNet still is relatively small – enables testing the whole set of instances, while manual evaluation is always laborious and can be performed only on a sample. Yet, the samples have been chosen as for the manual patterns (Israel, 1992), so the results can apply to the whole sets with a 95% confidence.

In both types of comparison we applied the standard measures of *precision* and *recall* (Manning and Schütze, 2001)[9].

Precision and recall are defined in the standard way: tp is the number of true positives (extracted pairs of LUs which are instances of the target relation), fp is the number of false positives (incorrect instances marked by algorithms as correct), fn – false negatives (correct instances in text but not extracted by the algorithm).

$$P = \frac{tp}{tp + fp} \tag{4.7}$$

$$R = \frac{tp}{tp + fn} \tag{4.8}$$

Note that the denominator in R accounts for correct patterns or instances that were either marked as incorrect or not extracted at all. We cannot treat the limited core plWordNet as the exhaustive description of relations. That is why recall in our approach only measures the *ratio of rediscovery* of the plWordNet structure. It is not a recall in terms of all correct instances in the corpus or patterns that the corpus supports.

Thus, following Pantel and Pennacchiotti (2006), we also use the *relative recall* measured in relation to the results of some other algorithm (Kurc, 2008, pp. 72):

$$R_{A|B} = \frac{R_A}{R_B} = \frac{\frac{tp_A}{C}}{\frac{tp_B}{C}} = \frac{tp_A}{tp_B} = \frac{P_A \times (tp_A + fp_A)}{P_B \times (tp_B + fp_B)} \tag{4.9}$$

where R_A and R_B denote the recall of the algorithms A and B, and C is the unknown number of instances occurring in the corpus.

We extracted a ranked list of possible instances which can be sorted in descending order by reliability. The values are real numbers and there is no characteristic point below which we can cut off the rest of pairs according to some analytical properties. Thus, instead of pure precision and recall, we prefer to use *cut-off precision* and *cut-off recall* calculated only in relation to some n first positions on the sorted list of results (instances or patterns).

In the end, then, we used three evaluation measures.

1. *Cut-off precision based on plWordNet* marks as correct only those instances and patterns that were found both in plWordNet and on an additional list provided *a*

[9] The F-measure could not be applied because of the limitations of recall based on plWordNet, to be discussed later.

priori by a human evaluator[10]. It is worth considering that the limited size of plWordNet can influence precision negatively. Some LUs are either not present yet or their synsets and all hypernymic links are incomplete. This precision is computed for each element on the list of instances.

2. *Precision based on human judgement* is evaluated according to a sample randomly drawn from the list of instances. Due to its cost, this evaluation measure was used only for selected experiments, see below. The error level of the sample was 3% and the confidence level was 95% (Israel, 1992).

3. *Recall based on plWordNet* is evaluated on the set of word pairs generated from plWordNet. This measure does not describe the recall in relation to the corpus used. In the case of many experiments, recall is also presented in the cut-off version.

The experiments were performed on exactly the same three corpora as for MSR extraction (Section 3.4.5):

1. IPIC (Przepiórkowski, 2004) with ≈ 254 million tokens,

2. the corpus of the electronic edition of a Polish newspaper *Rzeczpospolita* from January 1993 to March 2002 (Rzeczpospolita, 2008) with ≈ 113 million tokens,

3. a corpus of large texts in Polish collected from the Internet, with ≈ 214 million tokens.

In contrast with the experiments with the manually constructed patterns, there was no limit of the set of nominal LUs processed. Only the set of possible multiword LUs was predefined to accommodate the method of the recognition of multiword LUs based on the lexicon.

We tested several configurations:

ESP- – Espresso without generic patterns,

ESP-nm – Espresso without generic patterns, but with the extended reliability measure (4.6),

ESPmorf- – Espresso without generic patterns, but with additional morphological information encoding part of speech and values of selected grammatical categories: case (nouns), case and degree (adjectives), aspect (verbs),

[10] The list had resulted from the preliminary experiments and was next kept in use because of the limited size of plWordNet.

ESPfree- – extends **ESPmorf-** by the representation of the free order of the instance elements,

EST- — Estratto without generic patterns, exploiting specific features of Polish, e.g. the agreement on values of selected categories is represented,

EST-nm — Estratto without generic patterns, exploiting specific features of Polish language and the extended reliability measures (4.6),

EST+nm – the same as **EST-nm** but using generic patterns.

If not stated otherwise, the threshold for confidence is 1.0 for all ESP systems and 2.6 for EST systems. The number k of top patterns was set to $k = I + 2$, where I is the iteration count. There were four iterations. In those experiments whose results we present, the focus was only on the hyponymy/hypernymy relation. IPIC was selected as the main corpus, on which we ran all experiments with results presented in the tables.

We ran three groups of experiments on Espresso and Estratto (Kurc, 2008). We began with experiments designed to analyse the influence of the proposed extended reliability measure (4.6) and six pattern schemes: **ESP-**, **ESP-mn**, **ESPmorf-**, **ESPfree-**, **EST-nm** and **EST+nm**. The question was whether they improve the results, since they may better cope with certain characteristics of Polish. In Table 4.2, the precision based on human judgement is presented in the column labelled "Hum. eval.". The levels of precision defined in the column group labelled "Ranking" are achieved for the top subsets of instances described in the column group named "Inst.". They show on how large a portion of the extracted instances we can rely, and how strongly. Take, for example, the first row: **ESP-** extracted the top 8% of instances with the precision above 70% and the top 22% with the precision 60%. The higher the numbers, the higher concentration of positive instances in the upper part of the extracted list. The precision measured in relation to plWordNet is presented in the column "Prec. plWN" (the number of the extracted plWordNet instances is also given). The column labelled "Rel. R" refers to the recall calculated in relation to the result of **ESP-**.

The results of the first group of experiments, presented in Table 4.2, allow us to conclude that the modified reliability measure (4.6) performs better either in the case of the original Espresso scheme patterns (**ESP-nm** is the winner) or Estratto patterns which take into account some properties of Polish (**EST-nm** had the best overall result). The situation is less clear in the case of the precision based on plWordNet – the differences are smaller – but still **ESP-nm** and **EST-nm** produce better results than the other versions; plWordNet is relatively small, however, and this could bias the calculation. The manual evaluation showed that in fact plWordNet might be used only for a very rough estimation of precision. The plWordNet-based precision of **ESP-nm** versus **EST-nm** is almost identical, but **EST-nm** is much better in relation to the

	Hum. eval.	Ranking			Prec. plWN		Rel. R	Inst.
	[%]	70%	60%	50%	[%]	Inst.		
ESP-	39	8	22	43	36	501	1.0	3982
ESP-nm	47	5	14	62	37	561	1.54	6435
ESPmorf-	45	13	18	71	39	361	0.75	2600
ESPfree-	43	9	12	23	29	567	1.36	4621
EST-	54	10	27	–	30	651	1.71	4917
EST-nm	59	42	90	–	35	571	1.7	4457
EST+nm	37	18	32	52	27	1312	2.38	10000

Table 4.2: The influence of the extended reliability measure and changes in the pattern form (“Hum. eval.” – precision based on human judgement, “Ranking” – the number of the top instances above the precision threshold, “Prec. plWN” – precision in relation to plWordNet, “Rel. R” – relative recall relative to **ESP-**)

manual evaluation. It means that **EST-nm** starting from the same seeds acquired from plWordNet goes beyond the source and extracts many instances which are not described in plWordNet. This is a very promising feature concerning the potential application in expanding plWordNet.

We also observed that the value of the original reliability measure (4.1) decreases very fast. After the sixth iteration it goes far below 10-12. This explains the drop of the number of newly extracted instances. Applying the modified reliability formula (4.6) circumvents the problem.

Another matter of concern is the scheme of the patterns adjusted for Polish. It is clear that the application of the adjusted patterns produces better precision **EST-** and **EST-nm** in comparison to **ESP-** and **ESP-nm**. In the case of **EST+nm**, utilising the generic patterns, the precision is lower, but its relative recall shows its potential in extracting new instances. At the cost of reduced precision, the number of extracted instances increases by the factor 2.38 (the total number of the extracted instances depends on the number of instances above the threshold).

The second group of experiments was performed only for Estratto using generic patterns and the extended reliability measure, i.e. for **EST+nm**. The aim was to determine the influence of the algorithm parameters on the result. The following dependencies were investigated:

1. the influence of the confidence threshold on the precision of instances achieved within subsequent iterations,

2. the influence of the number of seeds on the induced patterns, and then the influence of the relation between instances and patterns induced by them,

3. the influence of the number of the top k patterns selected for the next iteration on the stability of the algorithm and the precision of instances,

4. the dependency on the filtering infrequent and very frequent patterns and instances.

5. the way in which various statistical similarity measures used in reliability calculation change the precision of the results.

	Human Eval. [%]	Relative Recall	Instances
EST+nm:th1.0	12	0.79	24552
EST+nm:th2.6	**37**	**1.00**	**10000**
EST+nm:th5.2	**48**	**0.54**	**4170**
EST+nm:5seeds	22	0.71	11882
EST+nm:10seeds	25	0.84	12476
EST+nm:15seeds	24	0.85	13189
EST+nm:5insts/1patt	24	0.83	12773
EST+nm:10insts/1patt	29	1.03	13188
EST+nm:40insts/1patt	37	1.00	10000
EST+nm:k4	37	1.00	10000
EST+nm:k8	**41**	**2.80**	**25361**
EST+nm:k12	38	2.70	26501

Table 4.3: The dependence of the algorithms on the parameter values (Kurc, 2008)

In case 1 it seems that the highest threshold gives the best results – see Table 4.3 and the first three rows, but a too high threshold decreases the total number of the extracted proper instances, as the relative recall is significantly decreased. There must, however, be a balanced ratio between instances selected for the next iteration and new patterns induced. With few instances, there is no statistical evidence to induce proper patterns, and EST/ESP crawls picking almost random patterns. That leads to the decrease in precision.

Initial seeds, case 2 (marked 'nnseeds' in Table 4.3, where nn is the number) are meant to generate a skeleton of a model of the lexico-semantic relation. If the number of seeds is not high enough, the best extracted patterns can be random. Of course, one could collect a small number of seeds that would indicate only expected patterns, but that would require a precise analysis of the corpus used for instance extraction. That is pointless, because by using more seeds one can acquire the same patterns with less effort.

The influence of the number of instances preserved between two subsequent iterations is similar to the influence of the number of seeds, see the rows marked 'nninsts/1patt' in Table 4.3 – nn preserved instances for one pattern. More instances kept, and next used for the evaluation of the patterns, give better description of the whole model. According to the experiments, at least 15 seeds and 10 instances for one pattern is required for stable behaviour of Estratto. The number of the preserved instances depend on the threshold, it cannot be too high.

Considering case 3, the results of the experiments presented in (Kurc and Piasecki, 2008) suggested that lowering k increases precision. This is only partially true. A more thorough evaluation performed in (Kurc, 2008), see the last three rows of Table 4.3, revealed a more complex picture. The value of k should be kept small due to the stability of a model in which a small group of elite patterns generates semantic relations. It should be too small either, in order to allow for the exploration of the instance space by a richer set of patterns. The result obtained for $k = 8$ is the best outcome produced by Estratto. The manually assessed precision is high in comparison to the other results and, at the same time, the relative recall is very high: many hypernymy instances have been extracted.

The data for case 4 are not presented in Table 4.3, but some experiments have shown that infrequent patterns (< 4 occurrences in IPIC) should be filtered out before generalization, because they introduce additional noise which causes good patterns to be evaluated as worse (Kurc, 2008, Kurc and Piasecki, 2008).

In case 5, the data show that PMI is much better than the z-score and t-score statistical measures of association (patterns versus instances) in the extraction of lexico-semantic relations. T-score results are especially disappointing. It might be due to the insufficient statistical evidence (the algorithm very often accepted instances which occur only once).

The third and the last group of experiments was prepared to check the ability of EST and ESP to use a different corpus and extract relations other than hyponymy/hypernymy.

The experiments showed that both EST and ESP can be successfully applied to different corpora. It seems, though, that each time the corpus is changed, a new confidence threshold must be somehow discovered. For IPIC, the threshold value was 2.6 but for the *Rzeczpospolita* corpus we have found 0.9 to work fine, cf (Kurc and Piasecki, 2008). Tests performed on the corpus of large documents from the Internet appeared to be unsuccessful. This is a rather special case, since most of the texts in this corpus are written in a literary style, so the language expressions are more complex. One should also expect fewer defining sentences than in utility texts. It seems that this kind of corpus requires more powerful patterns to catch syntactic dependencies. The precision of Estratto-based patterns on this corpus remains in an interesting contrast with the results for the manually created patterns (Table 4.1). Those results are better for the Internet-based corpus than for IPIC. We note that manual patterns use more expressive description of the morphosyntactic associations than Estratto patterns. The further development of Estratto should go in this direction. Presently, the manual patterns seem to be a valuable extension of the automatically extracted patterns.

The application of EST to different relation types appeared to be only partially successful. Tests on meronymy ended with the estimated precision below 30%, cf (Kurc, 2008). We see three main reasons of this failure. The expressive power of the patterns is too low and some important morpho-syntactic dependencies are missed. Meronymy is actually a set of quite varied sub-relations: it could be reasonable to try to extract each sub-relation separately. Finally, the trials ran on only one corpus.

On the other hand, initial experiments on extracting antonymy (only for adjectives) gave promising results. The human-judged cut-off precision reached 39%. Still, from the point of view of plWordNet expansion, extracting meronymy and antonymy requires further investigation.

Figure 4.5 shows examples of seeds and hypernymy instances extracted from IPIC by Estratto, version **EST+nm**. Here we list examples of patterns extracted from IPIC and used in the extraction of the instances that appear in Figure 4.5.

```
occ=31 rel=0.26803
(hypo:subst:nom) być    (hyper:subst:inst)
(hypo:subst:nom) is/are (hyper:subst:inst)

occ=20 rel=0.222222
(hypo:subst:nom) i inny    (hyper:subst:nom)
(hypo:subst:nom) and other (hyper:subst:nom)

occ=26 rel=0.103449
(hypo:subst:inst) a inny    (hyper:base:inst)
(hypo:subst:inst) but other (hyper:base:inst)

occ=15 rel=0.0684905
(hypo:subst:inst) przypominać (hyper:subst:acc)
(hypo:subst:inst) resemble    (hyper:subst:acc)

occ=41 rel=0.0263854
(hypo:subst:loc) i w inny    (hper:subst:loc)
(hypo:subst:loc) and in other (hper:subst:loc)

occ=86 rel=0.00708506
(hypo:subst:nom) stać się (hyper:subst:inst)
(hypo:subst:nom) become   (hyper:subst:inst)
```

```
occ=88 rel=0.0060688
(hypo:subst:acc) interp który być (hyper:subst:inst)
(hypo:subst:acc) interp which is  (hyper:subst:inst)
```

The plWordNet-related precision of the Espresso/Estratto algorithm is lower when measured on Polish corpora than the precision reported by Pantel and Pennacchiotti (2006). This might be due to a slightly different approach to precision evaluation, which was performed partially on the basis of the much smaller plWordNet. On the other hand, the results of the manual evaluation are similar to the results reported in (Pantel and Pennacchiotti, 2006). The results for different similarity measures based on reliability suggest that PMI gives the best results for the given test suite.

The adjustment of the pattern scheme to the characteristic features of Polish improved the precision over Espresso patterns using only word forms and parts of speech as features.

Estratto, the proposed modification of Espresso, succeeded in extracting hypernymy and antonymy from IPIC and the Rzeczpospolita corpus. Attempts to extract meronymy were unsuccessful. Meronymic pairs are present on the MSR-produced list of LUs the most semantically related to the given one, but with failure of the pattern-based attempts we do not have an additional source of knowledge to separate meronymic pairs from those lists.

We tested several parameters that have a significant influence on the Estratto algorithm. The most important of them appeared to be:

- the *number of seed instances*,
- the *confidence threshold*,
- the *number of the top k patterns* preserved between the subsequent iterations.

The number of seed instances should exceed 10. The confidence threshold strongly depends on the corpus; for example, for IPIC the best value found was about 2.6. Each time the algorithm is applied to a new corpus, both seed instances and the measure of confidence must be redefined. The number of the top k patterns should be low around 8. Such a number results in a more stable representation of the semantic relation. It is still unclear how to explore patterns that seem to be correct and are close to the top. Those patterns usually disappear in the next iterations, so some instances are also excluded from final results.

Espresso/Estratto is an intrinsically weakly supervised algorithm. That is true even though the preparation of an appropriate set of seeds leading to the extraction of patterns producing large and diverge set of extracted instances might require even some initial

Seed instances	
senator (*senator*)	mówca (*speaker*)
nazwa (*name*)	oznaczenie (*designation*)
Polska (*Poland*)	kraj (*country*)
Polska (*Poland*)	państwo (*state*)
wynagrodzenie (*remuneration*)	świadczenie (≈*benefit*)
agencja (*agency*)	jednostka (*unit*)
akademia (*academy*)	uczelnia (*university*)
alkohol (*alcohol*)	substancja (*substance*)
pożar (*fire (conflagration)*)	zdarzenie (*event*)
należność (*charge*)	zobowiązanie (*obligation*)
protokół (*protocol*)	dokument (*document*)
dolar (*dollar*)	waluta (*currency*)
broń (*weapon*)	przedmiot (*object*)
uposażenie (*salary*)	świadczenie (*benefit*)
obligacja (*bond*)	papier (*share*)
zapis (*record*)	dowód (*evidence*)
człowiek (*human*)	podmiot (*subject*)
żywica (*resin*)	spoiwo (*adhesive*)
Extracted instances	
szkoła (*school*)	instytucja (*institution*)
maszyna (*machine*)	urządzenie (*device*)
wychowawca (*tutor*)	pracownik (*employee*)
kombatant (*veteran*)	osoba (*person*)
bank (*bank*)	instytucja (*institution*)
pociąg (*train*)	pojazd (*vehicle*)
telewizja (*television*)	medium (*medium*)
prasa (*press*)	medium (*medium*)
szpital (*hospital*)	placówka (*institution*)
czynsz (*rent*)	opłata (*payment*)
grunt (*land*)	nieruchomość (*real estate*)
Wisła (*Wisła*)	rzeka (*river*)
świadectwo (*diploma*)	dokument (*document*)
opłata (*payment*)	należność (*charge*)
ryba (*fish*)	zwierzke (*animal*)
Włochy (*Italy*)	kraj (*country*)
jezioro (*lake*)	zbiornik (*reservoir*)
jarmark (*fair*)	impreza (*event*)
piwo (*beer*)	artykuł (*product*)
zasiłek (*dole*)	świadczenie (*benefit*)
powódź (*flood*)	klęska (*disaster*)
paszport (*passport*)	dokument (*document*)

Figure 4.5: Examples of hypernymy instances extracted by Estratto, version **EST+nm**

runs of Espresso/Estratto or browsing the corpus to find occurrences of promising LU pairs. A similar problem is with finding the appropriate parameter values. In our experience, trial runs of the algorithm for each corpus used are needed before getting results that satisfy our expectations.

Additionally it turned out that in order to maintain a stable representation of relations, there must be an appropriate ratio between patterns and instances. The pattern:instances ratio estimated during experiments is between 1:15 and 1:20. If there are fewer instances, the algorithm becomes unstable. Using more instances results in a longer computation time.

An interesting result is the observation of the "intensifying" patterns. Such patterns do not represent any particular semantic relation. When applied alone, they extract instances of relations of multiple types. When an intensifying pattern is combined with regular ones, it delivers additional statistical evidence to correct but infrequent instances. This lift the algorithm's precision. An example (Polish `w` means "in"):

```
(hypo/holo:subst:nom) w (hyper/mero:subst:inst)
```

We observed a problem with the number of instances collected by the **ESP+/EST+** versions of the algorithms that use generic patterns. This number is comparable to the number of instances extracted by **ESP-/EST-**, but one would expect it to be much higher. This might be a result of the characteristic features of the IPIC corpus or of the size of the validating corpus. This problem might be partially solved by using the Web as a validating corpus. Unfortunately, Polish LUs have multiple word forms, so Google queries must be more complicated. The other reason might be the limited expressive power of the patterns – an aspect of the algorithm that should be investigated.

The extended structure of Estratto patterns still seems to miss some lexico-semantic dependencies, especially in stylistically rich text. The experiments on extracting hypernymy from the Internet-based corpus, mostly consisting of literary texts, were unsuccessful. The first step towards strengthening patterns is to take into account possible agreements in elements of the patterns that match the instances. The patterns used in EST are very strict about grammatical categories. For example, the pattern

```
(hypo:subst:gen) i inny (hyper:subst:gen)
```

(two nouns in genitive) is treated as a completely different pattern from

```
(hypo:subst:inst) i inny (hyper:subst:inst)
```

(two nouns in instrumental).

It seems to be helpful to allow merging such patterns, maybe like this:

```
(hypo:subst:case1) i inny (hyper:subst:case1).
```

The results for **ESP-** and **EST-**, where there are no such strict constraints, suggest some increase in recall. Another way, much more complicated, is to enrich the pattern representation, so that additional syntactic information (at least about nominal LUs) can be used.

The list of acquired instances cannot be directly imported to plWordNet. First of all, the list is flat. There is no information on synsets. The percentage of erroneous LU pairs on the lists (such as 63% for **EST+nm**) is too high to trust the list as source of *automatic* expansion of the plWordNet hypernymy structure. Also, many positive LU pairs represent in fact quite remote hypernymic links.

These observations show the drawbacks, but there also are pluses. **EST+nm** extracted 3700 hypernymic LU pairs (37% of the 10000 LU pairs). This information can be combined with MSR$_G RWF$, producing higher values for wordnet relation instances. The MSR alone does not say what kind of relation made two LU closely semantically related. The information acquired by Estratto sheds light on this issue. Section 4.5.3 presents a fairly succesful algorithm based on this reasoning. A manual comparison of the LU pairs extracted by Estratto and the three manual patterns reveals that both sets are disjoint to some extent. We noted earlier that manual patterns are more expressive and can find hypernymic instances in language construction which are inaccessible for the present Estratto patterns. This can be changed in the future extensions of Estratto, but for now we used both types of patterns in the hybrid algorithm of plWordNet expansion in Section 4.5.3.

4.5 Hybrid Combinations: Patterns, Distributional Semantics and Classifiers

We noted at the end of Section 3.4.5 that Measures of Semantic Relatedness [MSRs] can recognize semantically related LUs with an accuracy approaching human performance. Still, MSRs produce lists of the k LUs most semantically related to the given LU x [MSRlist$_{(x,k)}$] with few instances of wordnet relations, and they do not know how to distinguish the direction of a relation. We named two ways of compensating for these drawbacks: introduce a classifier operating on MSRlists$_{(x,k)}$, capable of differencing relations, or combine a MSR with other sources of knowledge, including lexico-syntactic patterns or the existing wordnet structure. This subsection will examine both possibilities.

4.5.1 Classifiers for lexical-semantic relations

MSRs should extract wordnet relation instances well: recall is high (up to the limit of the size of the vocabulary of the underlying corpus) since any LU pair gets a value. Yet, high values do not tell us what kind of a relation links a given LU pair. We need to attach relation labels to the LU pairs related strongly enough. We also need to determine when two LUs are connected strongly enough to be in an wordnet relation. No threshold on the MSR values answer this question straight (Section 3.4.5). Now, to label LU pairs by relation labels or a catch-all *unrelated* label is a typical *classification problem*, for which Machine Learning is a tool of choice.

Separation into several classes is a harder classification task. Many algorithms work for two classes only or are better tuned for the two-way scenario. A wordnet's structure is fully defined by all its relations, but hypernymy is central, especially for nouns. Our first attempt is a classifier which assigns pairs of LU to the positive class *close hypernymy or near-synonymy* or to *other*. Experience of work on MSRs and pattern-based methods suggests that a finer-grained subdivision of the positive class is very hard.

Snow et al. (2005) proposed a supervised ML method of extracting hypernymy instances. They started from the idea of a supervised algorithm to combine a large number of lexico-syntactic patterns into a binary classifier of hypernymic LU pairs. The patterns had been extracted from a large corpus parsed by the dependency parser *MiniPar* (Lin, 1993). 752311 noun pairs $\langle n_i, n_j \rangle$ from PWN 2.0 at a distance no longer than four dependency links in the parse tree have been identified and classified as Known Hypernyms (14387) and Known Non-Hypernyms (737924, the ratio 1:50). This was based on the fact that n_j is an ancestor of the first sense of n_i in the PWN 2.0 hypernymy structure. Only "frequently-used" senses of each noun were taken into account.

Patterns were generated from classified noun pairs, as descriptions of the dependency paths that linked nouns in the occurrences. Such defined patterns are a slightly extended version of lexico-syntactic patterns in MSR construction (Section 3.4.2). Naïve Bayes and logistic regression algorithms were used to train classifiers on the collected data. Testing, done on noun pairs labelled in relation to PWN, was to distinguish non-hypernymic pairs from hypernym pairs at unrestricted distance. The best F-score in 10-fold cross validation was 0.348.

Next, Snow et al. (2005) combined a classifier of *coordinate nouns* (with a common hypernymic ancestor) with a hypernym classifier and other classifiers based on such sources as Wikipedia or PWN. In an evaluation on 5387 manually labelled noun pairs, 0.3268 was the best F-score for the corpus-based only models (without the use of structural information such as in the Wikipedia).

Kennedy (2006) analysed several modified versions of this method. The modifications concerned the data in the training/test corpus (varying the positive-to-negative pair ratio and the method of undersampling negative examples) and small differences in the way of formatting dependency paths. An additional classifiers based on a version of the *Supported Vector Machines* algorithm (Joachims, 2002) was applied too, achieving the best F-score 0.633 for a combination of a classifier and filtering based on Roget's Thesaurus.

Zhang et al. (2006) explored different types of syntactic dependencies at different levels of granularity in the construction of classifiers to find occurrences of relationships between named entities. Five main kinds of relationships with 24 different subtypes were considered. This approach is broadly similar but the different objective makes a comparison of the results difficult.

ML methods of extracting hypernymy pairs usually take lexico-syntactic features directly to build a classifier. Tens of thousands of features are typical, each carrying very sparse information. Most of such information "tells" the classifier about various aspects of semantic relatedness. Features that point to specific lexico-semantic relations are rare. Section 3.4.5 notes that near-synonyms and close hypernyms/hyponyms of an LU u would be expected close to the top of the list of LUs most semantically related to u, generated by a good MSR. An application of a syntactic analyser is also assumed: a deep parser in (Zhang et al., 2006) or a shallow dependency parser in (Snow et al., 2005, Kennedy, 2006). For many languages such tools are not available yet.

We propose to extract hypernymy pairs by relaxing both assumption. There are two phases (Piasecki et al., 2008):

1. extract the generic relation of semantic relatedness modelled by some MSR,
2. identify hypernymy instances – pairs of LUs – from the MSR's results.

The first phase can use all kinds of information that describes the semantics of LUs, depending on the MSR extraction method. The second phase concentrates on groups of semantically related LUs and applies specialised tests that distinguish specific lexico-semantic relations as subtypes of semantic relatedness. The tasks of the first phase are preliminary filtering and problem complexity reduction, so during the second phase a broader variety of ML methods can be used. An MSR of good accuracy can (by way of its high values) associate LUs that extremely rarely occur close by in the corpus at hand. Note that such occurrences are the precondition on any pattern-based method. MSRs condense information otherwise distributed among many lexico-syntactic patterns; in phase 2 we can concentrate on the most promising pairs.

The only assumption is the availability of a highly accurate MSR. During experiments we used an MSR based on the *Rank Weight Function* transformation [MSR$_{RWF}$], an earlier version of the *Generalised RWF* presented in Section 3.4.4. MSR$_{RWF}$ differs from MSR$_{GRWF}$ discussed earlier (Section 3.4) only in the transformation applied and in a slightly lower accuracy. A detailed presentation of the applied MSR$_{RWF}$ can be found in (Piasecki et al., 2007b, Broda et al., 2008).

The second phase begins with the extraction, for the given LU u, of a list S of k LUs most semantically related to u, denoted MSRlist$_{(u,k)}$. Any value of k will do, but we noticed that, for the MSR types we used, the percentage of LUs in a wordnet relation to u begins to deteriorate around $k = 20$. Next, we need a classifier to select a subset of S that includes near-synonyms and close hypernyms of u.

Instead of using frequencies of lexico-syntactic features collected from a corpus directly as attributes in learning the classifier, we want to identify a set of complex features that can give clues on the relation between two LUs. We intend to apply a kind of knowledge-based, partially linguistically-motivated, transformation of the initial feature space into a new space of reduced complexity: fewer features and maybe condensed information on the LU relations of interest. For a pair of LUs, the values of attributes are calculated prior to training or testing. This is done via co-incidence matrices constructed from large corpora. We generally work with the same matrices as in the MSR$_{RWF}$ construction.

In search for attributes, we drew on clues which can deliver information concerning the specificity of compared nouns, the extent to which they mutually share lexico-syntactic features, topic contexts in which they occur together and, last but not least, the value of their semantic relatedness. We now present the complete list of attributes used (a and b are noun LUs):

1. *semantic relatedness* $MSR(a, b)$ – the value returned by an MSR$_{RWF}$,
2. *co-ordination* – the frequency of a's and b's co-occurrence in the same coordinate noun phrase,
3. *modification by genitive* – the frequency of a's modification by b in the genitive form,
4. *genitive modifier* – the frequency of b's modification by a in the genitive form,
5. *precision of adjectival features* – the precision of repeating b's adjectival features by the set of a's features (for the calculation method, see formula 4.10 below),
6. *recall of adjectival features* – the recall of repeating b's adjectival features by the set of a's features (for details, see formula 4.11),
7. *precision of modification by genitive* – the precision of repeating b's features, which express modification by a specific noun in genitive, by the similar features of a (the calculation method is similar to that in formula 4.10),

8. *recall of modification by genitive* – the recall of repeating b's features, which express modification by a specific noun in genitive, by the similar features of a,

9. *global frequency* of a – the total frequency of a in the corpus,

10. *global frequency* of b – the total frequency of b in the corpus,

11. *number of significant adjectival features* of a – the number of adjectival features whose co-occurrence with a is statistically significant, e.g., according to the *t-score* measure,

12. *number of significant adjectival features* of b – the number of adjectival features whose co-occurrence with b is statistically significant,

13. *co-occurrence in text window* of a and b – the frequency of a and b co-occurring in the same wider text window, e.g., of the size $t_w = \pm 50$ tokens,

14. *significance of co-occurrence in text window* of a and b – the statistical significance of a and b co-occurring in the same text window, e.g., on the basis of the *t-score* measure,

15. *adjectival specificity* of a – after Caraballo (1999), calculated here (see formula 4.12) as the average number of adjectival features for a single occurrence of a in the corpus,

16. *adjectival specificity* of b – calculated according to formula 4.12,

17. *adjectival specificity ratio* – the ratio of a's adjectival specificity to b's adjectival specificity.

In subsequent discussion, we use the term *relevant LUs* jointly for near-synonyms, close hypernyms and close hyponyms that occur on MSRlist$_{(a,k)}$. From the point of view of the intended expansion of a wordnet, all three relations jointly mark potential placements of a in the hypernymy structure, so they may be relevant to the linguist's work.

We pass to the classifier only those LUs whose value of semantic relatedness is higher in comparison to other pairs of LUs, defined as by MSRlist$_{(a,k)}$ for some predefined k , but the exact value of MSR is still important. It is more likely that a relevant LU b will have a higher value of $MSR_{RWF}(a, b)$ – the attribute 1 – than non-relevant LUs. It is hard to find a global threshold for the MSR_{RWF} values guaranteeing some accuracy, but in the case of particular MSRlists$_{(a,k)}$ some characteristics points can be observed quite often. So we kept the MSR_{RWF} value as an attribute for a classifier to combine with the other information.

The next group of attributes is meant to tell the classifier directly about the possibility of hypernymy or co-hyponymy between a and b. The co-ordination attribute (2) is based on the lexico morpho-syntactic constraint `NcC` used for the MSR$_{RWF}$ extraction (Section 3.4.3, page 67). `NcC` looks for occurrences of syntactic co-ordination of a and b as constituents of the same composite noun phrase. It recognises only a limited set of conjunctions: *ani* (*neither, nor*), *albo* (*or*), *czy* (*whether*), *i* (*and*), *lub* (*or*), *oraz* (*and*). All these were manually identified as marking *semantic coordination* of the linked nominal LUs, possibly indirect co-hyponyms ("*coordinated terms*" in (Snow et al., 2005)).

During matrix construction, occurrences of `NcC` are recorded for a LU x with all nominal lexical elements, very often more than 100000. Here, we pay attention only to nominal LUs in the classified pairs – potentially all LUs described by the given MSR. The value of (2) is the frequency with which the constraint is met for a and b co-occurring in the same sentence.[11] We assumed that *co-ordination* is more frequent for potential co-hyponyms and hypernyms.

A manual investigation of instance pairs of hypernyms in the IPI PAN Corpus of Polish[12] [IPIC] (Przepiórkowski, 2004) showed that, surprisingly, they often occur as the noun phrase head and its noun modifier in the genitive case. Even more frequent is meronymy expressed by the genitive modification. The classifier receives information on the frequency of this syntactic relation in both directions, when a is modified (3) and is the modifier (4). Both attributes are based on the same lexico-morphosyntactic constraint `NmgC` used for MSR extraction, presented in Figure 3.6 and discussed in Section 3.4.3. `NmgC` is based more on the relative positions of both nominal LUs than on agreement. Just as attribute (2), `NmgC` was used only to detect associations between LUs described by the MSR.

The idea of the precision of repeating b's features by a's features, used in attributes 5 and 7, is modelled after the MSR in (Weeds and Weir, 2005). We want to analyse the *additive precision* with which, using a's features, we refer to ("retrieve") b's features. The precision is defined as follows:

$$P^{add}(a,b) = \frac{\sum_{i \in F(a) \cap F(b)} \mathbf{M}[a,i]}{\sum_{j \in F(a)} \mathbf{M}[a,j]} \tag{4.10}$$

- $F(x)$ is the set of features occurring frequently enough with x, according to a test of statistical significance, e.g., a *t-score* test,

- $\mathbf{M}$ is a co-incidence matrix that represents the given set of features; for attribute

[11] The corpus is processed with the granularity of sentences – identified by a simple sentencer.

[12] Especially in the part called HC (Section 4.1) – sentences that contain pairs of known hypernyms. HC has been extracted to facilitate manual construction of lexico-syntactic patterns.

5 the matrix of adjectives and adjectival participles $\mathbf{M}_{adj}$ is used, while for attribute 7 it is the matrix $\mathbf{M}_{Ng}$ of modification by nouns in the genitive case.

The additive recall of repeating b's features a's features, used in 6 and 8, is calculated similarly to P^{add} (Weeds and Weir, 2005):

$$R^{add}(a,b) = \frac{\sum_{i \in F(a) \cap F(b)} \mathbf{M}[b,i]}{\sum_{j \in F(b)} \mathbf{M}[b,j]}, \tag{4.11}$$

Additive precision and recall are calculated for each type of descriptive features separately, but the four attributes together are intended to show to what extent the description of a is included in the description of b. We assume that the possible descriptions of a hyponym are covered by the possible descriptions of its hypernym. Precision and recall allow us to test this dependency in both ways and measure its strength.

During the preliminary experiments, we noticed that nouns semantically related by situation type are difficult to distinguish from relevant nouns. In order to capture the difference, we added two attributes intended to signal a kind of topic similarity – the two nouns would be used in the description of the same topics. That is why the value of attribute 13 is the frequency of co-occurrence of a and b in a quite large context of ± 50. There also are no restrictions on these contexts. We want to record any co-occurrence. In attribute 14 this information is filtered and emphasised by the *t-score* test. However, we tested both versions as elements of a training/test vector.

With the next group of features, we try to describe how specific both nouns are and to get some information on the relation of hypernymy levels of a and b. First, the global frequency of a noun can say something about its generality – the attributes 9 and 10. It has not been normalised, but in all experiments the same corpora were used. Second, we also test the number of different significant adjectival features of both nouns – the attributes 11 and 12. We expected that hypernyms were modified by the larger number of adjectival features. Finally we apply to the description of both nouns a measure of adjectival specificity (15 and 16) following the proposal in (Caraballo, 1999) (a similar measure was proposed by Ryu and Choi (2006)):

$$spec(a) = \frac{\sum_i \mathbf{M}_{adj}[a,i]}{globalTf(a)} \tag{4.12}$$

$\mathbf{M}_{adj}$ is the co-incidence matrix with adjectives and adjectival participles, and $globalTf(a)$ is the total frequency of a in the corpus, that is to say, attribute 9.

Some machine learning methods (C4.5, for example) would find it difficult to get the ratio of both specificity measures, so we explicitly added this ratio (17) to the attribute set.

4.5.2 Benefits of classifier-based filtering for wordnet expansion

The MSR for the experiments and the values of all attributes were generated from two corpora combined – both were used in other our experiments. Their more detailed description can be found in Section 3.4.5. One was IPIC with $\approx$ 254 million token. The other was the corpus of the daily *Rzeczpospolita* with $\approx$ 116 million token (Rzeczpospolita, 2008).

MSR$_{RWF}$ was the same as that proposed by Piasecki et al. (2007b). Its construction was based only on two types of lexico-morphosyntactic constraints: modification by a *specific adjective or adjectival participle* (`AdjC` in Section 3.4.3, page 67), and coordination with a *specific noun* (`NcC`).

All nouns, adjectives and adjectival participles from the combined corpora were used accordingly as the lexical elements of constraint instances. MSR$_{RWF}$ provided a description of 13298 nominal LUs and achieved the accuracy of almost 91% in WBST+H, see Section 3.3.1 generated from the plWordNet version June 2008.

We used plWordNet as the main source of training/test examples. Following the main line of the experimental paradigm of (Snow et al., 2005), we generated from plWordNet two sets of LU pairs: Known Hypernyms [KH] and Known Non-Hypernyms [NH]. Our goal is to support linguists by presenting relevant pairs of LUs. Similarly to (Snow et al., 2005) we constructed the set of Known Hypernyms from LU pairs $\langle a, b \rangle$ where b is a direct hypernyms of a or a hypernymic ancestor of a. In contrast with (Snow et al., 2005), we allowed only for the limited hypernymic distances in all KH sets. Aiming at a tool to support linguists, we did not want remote associations among classified positively LU pairs.

Hypernymy path length guided experiments with two different divisions of the two groups. We wanted to investigate to what degree we can distinguish closer and more remote hypernyms. We generated four data sets from the plWordNet version April 2008:

H the set of pairs: direct hypernym/hyponym (2967 pairs) – in all experiments **H** was included in KH,

P2 pairs of LUs connected by the path of the two arcs in the hypernymy graph – **P2** was included in KH (2060 pairs),

P3 pairs of LUs connected by a path of three or more hypernymy arcs, in NH (1176 pairs),

R pairs of words randomly selected from plWordNet in such way that no direct hypernymy path connects them, NH (55366 pairs, including co-hyponyms).

After initial experiments, we noticed that the border space between typical elements of KH and NH is not populated well enough, especially considering its importance for Machine Learning. We manually annotated randomly selected pairs of LUs which occurred on MSRlists$_{(a,20)}$ for the LUs described by the MSR.

From this selection, 1159 pairs classified as not relevant were collected into a set **E**. In some experiments, we added **E** to NH, see below.

We experimented with two training sets produced by combining our regular data sets. Test sets were excluded randomly from training sets during tenfold cross-validation. Training sets are named in Table 4.4 according to the following description scheme:

$$KH_1 + \ldots + KH_n, NH_1 + \ldots + NH_m$$

i.e. first the sets comprising KH are listed, next the sets from NH. The training set H+P2,P3+R includes only pairs extracted from plWordNet. It consists of 5027 KH pairs (H+P2) and 56531 NH pairs (P3+R). Tests on this set were done only on data already present in plWordNet. It is also more difficult than the sets used in (Snow et al., 2005), because the classifier is expected to distinguish between close hypernyms and more indirect hypernymic ancestors (P3 included in NH).

Because plWordNet (the version June 2008) was still small, the second training set was extended with the set **E** of manually classified pairs. We added only negative pairs, assuming that positive examples are well represented by pairs from plWordNet, while more difficult negative examples are hidden in the huge number of negative examples automatically extracted from plWordNet. The second training set consists of 5027 KH (H+P2) and 57690 NH (P3+R+E).

In the experiments, we used Naïve Bayes (Mitchell, 1997) and two types of decision trees, C4.5 (Quinlan, 1986) and Logistic Model Tree [LMT] (Landwehr et al., 2003), all in the versions implemented in the Weka system (Witten and Frank, 2005). Naïve Bayes classifiers are probabilistic, C4.5 is rule-based, and LMT combines rule-based structure of a decision tree with logistic regression in leaves. In order to facilitate a comparison of classifiers, we performed all experiments on the same training-test data set. Because we selected C4.5 as our primary classifier, and we generated examples from the same corpus (so the frequencies occurring as values of some attributes could be compared directly), we did not introduce any data normalisation or discretisation. The range of data variety was also limited by the corpus used. The application of the same data to the training of a Naïve Bayes classifier resulted in a bias towards its more memory-based-like behaviour. According to the clear distinctions in the main group of the applied data sets, however, the achieved result was positive, see Table 4.4.

All experiments were run in the Weka environment (Witten and Frank, 2005). In each case, we applied tenfold cross-validation; the average results appear in Table 4.4.

Because some classifiers, for example C4.5, are known to be sensitive to the biased proportion of training examples for different classes (here, only two), we also tested the application of random subsampling of the negative examples (NH) in the training data. The ratio KH:NH in the original sets is around 1:10. In some experiments the ratio was randomly reduced to 1:1 (the uniform distribution of probability was applied in drawing a new subset NH).

	P	R	F_1	P	R	F_1
ratio	1:1			1:10		
Naïve Bayes						
H+P2,P3+R	89.80	47.10	61.79	46.30	45.80	46.05
H+P2,P3+R+E	84.70	59.10	69.62	34.60	53.50	42.02
C4.5						
H+P2,P3+R	82.10	77.50	79.73	66.90	43.10	52.43
H+P2,P3+R+E	81.70	78.40	80.02	60.70	39.90	48.15
LMT						
H+P2,P3+R	81.80	80.60	81.20	72.80	39.40	51.13
H+P2,P3+R+E	81.00	78.20	79.58	65.40	34.50	45.17

Table 4.4: Evaluation for both sets using tenfold cross-validation

Precision and recall are calculated in Table 4.4 according to the description of examples extracted from plWordNet (H, P2, P3, R) or defined manually (E). The results achieved by both decision trees are very similar, and high by all three measures. However, the inclusion of the set E decreases the result significantly in comparison to the high ratio $|R| : |E|$, that is to say, a small number of more difficult examples negatively influence the result. The R set includes more obvious and on average less closely semantically related pairs of LUs; it is generated randomly from plWordNet, but E includes only tricky examples. That is why we ran additional tests on a separate set of LU pairs selected randomly from MSRlists$(a, 20)$ generated using MSR$_{RWF}$. The set was annotated manually, and will be referred to as the *manual test set* (M). The best classifiers shown in Table 4.4 appeared to be biased towards positive decision, contrary to the classifiers trained on the 1:10 version of the learning data.

In Figure 4.6 we present sample results of the classification selected from one of the folds of the tenfold cross-validation (classifier C4.5, ratio KH to NH 1:10, **E** included in NH)[13].

We prepared the set M in order to go outside plWordNet with the tests and to look into the work of the classifiers from the point of view of their potential application in linguistic practice. As we wrote earlier, the set M was selected randomly from pairs of LUs with the highest value of semantic relatedness according to MSR$_{RWF}$. M consists of 1984 negative and 316 positive examples.

The C4.5 classifier trained on the KH=H+P2 and NH=P3+R+E with the ratio 1:10 achieved a 21.69% precision, a 50.32% recall and a 30.31% F-score. This is a little lower than the best result achieved by (Snow et al., 2005) using corpus-based attributes

[13] Many words in these pairs are polysemous in both languages. The English translations “select” the intended meaning.

True positives	
akt (*act*)	ustawa (*bill*)
bank (*bank*)	firma (*firm*)
emocja (*emotion*)	smutek (*sadness*)
intelekt (*intellect*)	przymiot (*attribute*)
licencja (*licence*)	zezwolenie (*permission*)
pragnienie (*desire*)	ochota (*willingness*)
terytorium (*territory*)	kolonia (*colony*)
warzywo (*vegetable*)	kartofel (*potato*)
False positives	
celnik (*customs officer*)	policja (*police*)
czynsz (*rent*)	oprocentowanie (*interest*)
dochód (*income*)	dotacja (*donation*)
nonszalancja (*nonchalance*)	rozrzutność (*profligacy*)
odpad (*waste*)	produkt (*product*)
problem (*problem*)	rodzina (*family*)
temat (*topic*)	dostarczyciel (*provider*)
zachwyt (*admiration*)	zdumienie (*astonishment*)
True negatives	
człowieczeństwo (*humanity*)	prorok (*prophet*)
licencja (*licence*)	zarządzenie (*regulation*)
opis (*description*)	hipoteza (*hypothesis*)
ślub (*wedding*)	kochanek (*lover*)
tempo (*speed*)	sport (*sport*)
trybunał (*tribune*)	sejm (*diet (parliament)*)
False negatives	
linia (*line*)	ogonek (*queue*)
konstrukcja (*construction*)	twierdza (*fortress*)
nieprzychylność (*unfriendly attitude*)	emocja (*emotion*)
podpora (*support*)	kula (*sphere*)
zakochanie (*infatuation*)	emocja (*emotion*)

Figure 4.6: Example results of the classification of LU pairs acquired from plWordNet as relevant (near-synonyms and close hypernyms) and not relevant. The classifier was C4.5, the positive to negative ratio 1:10; manually prepared negative examples were used together with automatically generated examples

only (F-score = 0.3268) and measured in relation to manually annotate examples. Our problem setting is more difficult (we expect the classifier to distinguish e.g. between P2 and P3, while Snow *et al.* included all indirect hypernyms in KH) and we had much fewer learning examples. Also, Snow *et al.* worked with a hybrid system that combined the hypernymy classifier with a MSR. It is more related to our WordNet Weaver system presented in the next section. Snow *et al.* had the best F-score of 0.2714 for the classifier-only version.

Leaving aside automatic evaluation, one can notice that the percentage of false positives is still significantly below 50%, which is a ratio that seems acceptable for a tool to support linguists. On the other hand, the number of LU pairs presented to linguists dropped dramatically in comparison to MSR_{RWF} alone, from 2300 to 733 – 31.87% of the initial list. The classifier cannot be used alone as a support tool, but its ability 'concentrate' KH pairs in the positively classified group will be leveraged in the next section for the construction of a tool combining different types of evidence in expanding plWordNet.

The results achieved on M for all classifiers were much poorer than the results on sets selected from plWordNet. We tried SVM as well, hoping for its usually good performance on numerical features without discretisation, but in contrast with the findings of Kennedy (2006) we have not achieved any valuable result.

In Figure 4.7 we present examples of classifier decisions made for elements of set M (classifier C4.5, ratio KH to NK 1:10, **E** included in NK).

A manual inspection of false positives in the classification results on set M shows that many pairs are co-hyponyms. They can be treated as positive answer from a linguists's point of view, but we tried to train the classifier not to select co-hyponyms as relevant pairs.

The results achieved on the data extracted from plWordNet are very promising, especially when we compare them to the results of similar experiments in (Snow et al., 2005), where the highest value of F-score was 0.348. A direct comparison, however, is not possible, because we used examples of KH and NH generated directly from plWordNet, not from sentences in the corpora. Randomly generated pairs can include a larger percentage of obviously negative cases. On the other hand, plWordNet is much smaller than PWN applied in (Snow et al., 2005), so some NH pairs are in fact relevant pairs not yet added to plWordNet. This introduces substantial noise during training.

The results on the manually annotated set M, and manually inspected, show that the performance of the classifiers on real data is lower. They have problems with distinguishing co-hyponym pairs from relevant pairs, and there are more errors for less obvious cases. Still, if we consider a task of delivering valuable suggestions to the linguists, we have achieved an enormous improvement in comparison with the lists of k most semantically related LUs. That is to say, a majority of the list elements are

True positives	
akredytacja (*accreditation*)	zezwolenie (*permission*)
anegdota (*anecdote*)	opowieść (*tale*)
dwója (*bad (lowest) mark*)	dwójka (*dyad, pair*)
forteca (*fortress*)	budowla (*edifice*)
forteca (*fortress*)	zamek (*castle*)
incydent (*incident*)	zajście (*incident*)
instrument (*instrument*)	przyrząd (*example*)
owca (*sheep*)	jagnię (*lamb*)
False positives	
abonent (*subscriber*)	odbiornik (*receiver*)
cmentarz (*cemetary*)	zakwaterowanie (*quarters*)
chwilka (*fleeting moment*)	berbeć (*toddler*)
gniew (*anger*)	strach (*fear*)
jesion (*ash tree*)	konar (*bough*)
owoc (*fruit*)	grzyb (*mushroom*)
palec (*finger, digit*)	nos (*nose*)
paliwo (*fuel*)	odpad (*waste*)
True negatives	
aktyw (*activists*)	przychód (*income*)
kompletność (*completeness*)	zgodność (*consistence, concordance*)
oś (*axle*)	kierunek (*direction*)
otyłość (*obesity*)	nowotwór (*cancer*)
ożywienie (*animation*)	postęp (*progress*)
False negatives	
agenda (*agenda*)	przedstawicielstwo (*diplomatic agency*)
alergia (*allergy*)	patologia (*pathology*)
ankieta (*survey*)	badanie (*investigation*)
komisariat (*police station*)	urząd (*office*)
lądowanie (*loading*)	manewr (*maneuver*)

Figure 4.7: Example results of the classification of LU pairs not present in plWordNet (at the time of the test) as relevant (near-synonyms and close hypernyms) and not relevant. The classifier was C4.5, the positive to negative ratio 1:10; manually prepared negative examples were used together with automatically generated examples

eliminated, but the error of elimination is small. Even so, we are still rather far from a support tool truly valuable from the linguists' point of view.

Because of the small size of plWordNet, it will be a laborious process to prepare a more demanding training set. In the case of each LU pair we can suspect that it is not yet described in plWordNet – building the set means expanding the wordnet. Nonetheless, it can be done and some bootstrapping approach can be applied in improving the classifier and expanding the wordnet. The next section presents work along these lines.

In contrast with (Snow et al., 2005), who use directly lexico-syntactic features, we proposed a two-step approach. It is intrinsically based on MSR, on whose quality it somewhat depends. On the other hand, a good MSR can introduce a general description of relations among LUs and deliver knowledge derived from a very large number of contexts, not only direct LU co-occurrences. The complex attributes designed for the classifiers are a form of pre-processing. They express condensed information that facilitates the classifiers' decision processes. The results achieved on the manual test set M shows that the present set of attributes does not give enough evidence for distinguishing near-synonyms and close hypernyms from co-hyponyms. More research is necessary on other possible sources of knowledge.

4.5.3 Multicriteria voting in wordnet expansion

A wordnet is built of LUs, synsets and relation links. After a rather unsuccessful attempt to acquire lemmas for LUs from corpora (Section 2.4), we took an initial batch from a small dictionary (Piotrowski and Saloni, 1999). We tackled the extraction of wordnet relation instances several times. We considered Measures of Semantic Relatedness [MSRs] (Section 3.4), manually constructed patterns (Section 4.1), automatically extracted patterns (Section 4.3) and a classifier-based method (Section 4.5.1). We have not achieved results better than around 30% of accuracy, but many symptoms suggest that a combination of algorithms can improve the accuracy a lot. In this section we will investigate thoroughly this possibility. The extraction of synsets, on the other hand, seems to be a serious problem. We could notice this in Section 3.5. The best clustering algorithm produces interesting results,but is still far from being a source of automatically extracted synsets. Clustering of LUs is a self-organising process and therefore raises expectations which in our case have not been met. From the point of view of future wordnet user one would expect, if not directly synsets, than some general but intuitively distinguished and useful classes as represented by the higher levels of the hypernymy structure.

Clustering algorithms also tend to produce a flat set of clusters. Changing such a set of clusters into a hierarchy poses two problems: how to identify the right shape of the tree and how to label higher levels of the cluster tree with the adequately general LUs. Moreover, most hierarchical clustering methods produces strict trees, while a wordnet hypernymy structure is normally a graph. In any case, no automatic method can come up with a credible top portion of a wordnet hierarchy. That is why the top levels of plWordNet's hypernymy hierarchy have been built manually, and we defined the focus of our research as a semi-automatic expansion of the core plWordNet. The constructed core plWordNet then serves as the springboard for what have turned out to be useful suggestions of attaching new lemmas to particular synsets in plWordNet. Such lemmas would be attached as related to, but not necessarily synonymous with, LUs in those synsets.

Several projects have explored the idea of building an expanded wordnet over an existing one. Most of them are focused, and have been tested, only on PWN. The advantage is the possibility of using the wordnet structure already in place, especially the hypernymy structure, as a knowledge source.

Caraballo (1999, 2001) discusses an interesting attempt to overcome those problems. In her approach, the meaning of nouns is described simultaneously in two ways. In a distributional semantics model, for each noun a vector is constructed with the co-occurrence frequencies of this noun and other nouns in coordinate and appositive constructions. The frequencies are collected from parsed text. In a pattern-based model, hypernym pairs are extracted by Hearst's pattern (Hearst, 1992) `X, Y, and other Zs`. The vectors give a cosine-measure similarity of nouns and noun clusters. A binary tree of clusters is built following the scheme of agglomerative clustering. Next, internal tree nodes are assigned hypernyms of the branches by extracting from the pattern-based pairs the most frequent hypernyms of the LUs in the given branch. Finally, the binary tree is "compressed" by removing internal nodes that have no hypernyms assigned or represent the same hypernyms as their parent node. A manual evaluation of a randomly selected sample showed that on average 33% of nouns were assigned correctly as hyponyms of the examined hypernyms. The sample was very small and not representative, and a 33% precision is similar to the precission achieved in our experiments on pattern-based hypernym extraction. Carraballo's approach, while interesting, required parsing (a drawback if no good parser is available) and was applied to a limited domain of economy and texts from the Wall Street Journal. The achieved precision seems limited and directly correlated with the precision of the patterns, and the constructed hierarchy is far from the wordnet synset structure: the number of internal nodes is small in comparison to the number of leaf clusters and their large size.

Alfonseca and Manandhar (2002) assigned to synsets a meaning representation based on distributional semantics model, and treated the hypernymy structure labelled in that way as a kind of decision tree. To find a site for a new lemma, the tree is traversed top-down each time, choosing a branch with the highest distributional similarity. The top-level synsets were mostly very general, so they introduced a limited propagation of meaning vectors from children to parents.

Witschel (2005) applied a more radical decision-tree model with recursive upward propagation of meaning descriptions. The propagation only stops in the root, and the description of the upper nodes represents the description of descendants. A synset's *semantic description* is a set of LUs most similar to LUs from this synset. Similarity calculation, following the distributional semantics model, is based on co-occurences of LUs in corpus. Semantic descriptions of children nodes are recursively propagated to parents and merged with their initial description. The resulting tree of semantic descriptions is then used as a decision tree to assign new lemmas. We select a branch by the highest similarity with a new lemma measured by the degree of matching between descriptions. Downward traversal stops in a node in which the mean of the similarity values with branches is greater than their variance. Evaluation was performed only on two subtrees taken from GermaNet: *Moebel* (*furniture*) (144 children) and *Bauwerk* (*building*) (902 children). The best accuracy of the exact classification was 14% and 11% respectively, comparable to that achieved by Alfonseca and Manandhar (2002).

Widdows (2003) represented LU meaning by the set of *semantic neighbours* – k most similar LUs. The main idea for attaching a new lemma was to find a site in the hypernymy structure in which its semantic neighbours are concentrated. For semantic similarity calculation, each LU was first described by the co-occurrence, in a 15-word text window, with the selected 1000 most frequent one-word LUs. Parts of speech were attached to words in the experiments that gave the best results. Similarity values were computed as in the Latent Semantic Analysis algorithm (Landauer and Dumais, 1997), cf Section 3.4.2. For the given LU and its first k semantic neighbours, a hypernym h is chosen as its *label* (attachment point), such that it gives the highest sum over *affinity scores* between the subsequent neighbours and h. The affinity score is negative for neighbours which are not hyponyms of h, and positive otherwise, with higher value for neighbours closer to h.

Evaluation was on the British National Corpus (BNC, 2007) and randomly selected common nouns, 200 each from three frequency ranges: $>$1000, $[500, 1000]$ and $<$500. During experiments, sites identified by the algorithm were compared with their exact hypernyms. It is unclear how many labels were taken into account, one or four. Widdows (2003) writes:

> Since our class-labelling algorithm gives a ranked list of possible hypernyms, some credit was given for correct classifications in the top 4 places.

The best accuracy of finding the direct hypernym among 4 highest ranked labels was 15% for 3 neighbours, but the overall classification (considering hypernyms up to 10 links away) gave only 42.63%. The best accuracy of the overall classification was 82.06% but the accuracy of the exact placing was reduced to 10.15%

Snow et al. (2006) cast the expansion of wordnet hypernymy structure in terms of a probabilistic model. Attachment of new elements of the structure transforms the former structure **T** into a new structure **T**'. Among many possible **T**', the most appropriate one is probably the one that maximises the probability of the introduced change in relation to the evidence at hand. The change caused by the addition of one new relation instance R_{ij} is described in (Snow et al., 2006) as follows:

$$\Delta_T(R_{ij}) = \frac{P\left(\dfrac{\mathbf{E}}{\mathbf{T'}}\right)}{P\left(\dfrac{\mathbf{E}}{\mathbf{T'}}\right)} \tag{4.13}$$

T and **T**' are the old and new *taxonomies* (hypernymy structures), the latter resulting from adding the R_{ij} instance of hypernymy. **E** is collected evidence (of any kind).

The computation of the complete *multiplicative change* is based on all added relation instances, as well as *implied* relation instances. For example, adding a new hyponym to the LU y implies the hypernymy relation to all hypernymic ancestors of y. The algorithm of taxonomy extension proposed by Snow et al. (2006) works according to the best-first search scheme that maximises a criterion based on the multiplicative change calculated for the extended and old taxonomies. The sources of evidence applied during experiments with expanding PWN were:

- a classifier-based algorithm of extracting hypernymic LU pairs on the basis of lexico-syntactic relations (proposed in (Snow et al., 2005) and briefly discussed in Section 4.5.1),
- a proposed algorithm of extraction of (m, n)-*cousins* derived from the algorithm presented by Ravichandran et al. (2002).

The relation of (m, n)-cousins holds for those noun pairs which have a common hypernymic ancestor at a distance of, respectively, m and n. The algorithm of extraction of (m, n)-cousins is based on a two-step procedure. First, nouns occurring in 70 million webpages are clustered into 1000 clusters. For each noun pair, the similarity calculation is based on shared clusters and the minimum across cosine measure between the nouns and cluster centroids. Second, a classifier (based on softmax regression) classifying noun pairs as cousins is trained on cousins extracted from PWN and described by their similarity in relation to the cluster-derived similarity.

Both classifiers used by Snow et al. (2006), i.e. classifier of hypernymic pairs and (m, n)-cousins return probabilities of their decisions. For new nouns (not present in PWN), the decisions suggest possible hypernyms and cousins in the PWN hypernymy structure and probabilities returned by the classifiers are used in computing multiplicative changes and identifying the hypernymy links finally added by the algorithm.

Snow et al. (2006) evaluated manually the first n automatically added hypernymic links. Because n was up to 20000 in the last experiment, only randomly selected samples were assessed. The applied uniform size of the samples equal to 100 for $n > 1000$ was too small to ascribe the results of the evaluation to the whole sets with sufficient statistical confidence. Among different types of evaluation performed, the *fine-grained* one seems to be the most interesting from our point of view. The evaluators were asked: "is X a Y?", where $\langle X, Y \rangle$ is an added link. It is not clear in (Snow et al., 2006) whether only direct hyponym/hypernyms counted as positive. For each pair of nouns $\langle i, j \rangle$, where i is unknown, the algorithm finally selects only one sense of j, so only the best hit is added or evaluated. According to this setting of fine-grained evaluation, the achieved precision of 84% for $n = 10000$ is high, but may be hard to compare with other approaches, including ours (to follow soon), because it is given only for the best hit and the basic criterion (cited above) is not precise.

Alfonseca and Manandhar (2002) and Witschel (2005) use only one knowledge source and work locally on the hypernymy tree. Each decision is based on the properties of the currently processed node. Widdows (2003) considers a broader context of several points in the structure but also uses only one type of evidence. Caraballo (1999) combined two types of information, patterns and MSR, but the extracted structure seems to be too far from the proper hypernymy structure. Snow et al. (2006) combine two knowledge sources and utilise not only vertical structure but also horizontal structure of cousins (direct and indirect co-hyponyms). The assumption was that the results of all classifiers can be described probabilistically – not the case for lexico-syntactic patterns. One of their classifiers is based on processing a corpus by parsing, a step not feasible for many natural languages.

How To Combine Extraction Algorithms

In the previous sections we have reported on several methods of extracting lexico-semantic relations [LSR] for Polish. None of them individually has reached the accuracy level required in a support tool for linguists. We will now investigate combinations of the following methods:

- a measure of semantic relatedness based on the Rank Weight Function, written as MSR_{RWF}, developed for Polish nouns and presented in (Piasecki et al., 2007a) – MSR_{RWF} extracts closely semantically related LUs with high accuracy, but the extracted LU pairs belong to a range of LSRs, not only to the typical wordnet relations;
- post-filtering LU pairs produced by MSR_{RWF} with a classifier presented in the Section 4.5.1 called here C_H – the percentage of LSR instances increases among the filtered pairs;

- three manually constructed lexico-morphosyntactic patterns presented in Section 4.1: *JestInst*, *NomToNom* and *mIInne*;
- the results of the *Estratto* algorithm application (Kurc, 2008, Kurc and Piasecki, 2008) discussed in Section 4.3 – the accuracy is higher than for manual patterns, but both types of patterns seem to complement each other somewhat.

The accuracy of all methods of distinguishing pairs of synonyms, close hypernyms and meronyms is at most around 30%. This is too low to support lexicographic work effectively. We note, however, that all three patterns and *Estratto* produce lists which share a limited number of pairs and that shared pairs significantly gain accuracy, more as the number of lists increases. A very similar correlation was observed in comparing C_H with the patterns. All analysed methods use slightly different markers of LSR in text and explore different pieces of information in text. We assumed that by combining the results of different methods we could achieve better accuracy in extracting LU pairs interesting for a linguist who is adding material to a wordnet.

One problem remains. Combined methods may better differentiate pairs of synonyms, hyponyms/hypernyms and meronyms/holonyms from LUs related otherwise, but still have lower accuracy in differentiating among the three. When processing a new lemma, however, all sites of its attachment to the wordnet structure are almost equally important – all three relations constrain the LU meaning. We assume that a method of robust extraction of LU pairs in either of the three relations would make an interesting tool for linguists.

Moreover, if all extraction methods constructed so far have significant problems with differentiating between near-synonyms, close hyponyms/hypernyms and meronyms, one should not expect any wordnet expansion based on those methods to differentiate between the three. Starting with this assumption, we proposed a method of *activation-area attachment*: a new lemma is automatically attached to a small area in the wordnet hypernymy graph rather than to one synset.

This method has been inspired by the general idea of learning in *Kohonen networks* (Kohonen, 1982). In a Kohonen network, a new learning example is used to modify not only the most similar neuron – the *winner* – but also neurons located in a small distance from it. The further the given neuron is from the winner, the smaller is its change caused by this learning example. The distance is measured by the number of links in the graph structure of the neural network (mostly organised in two-dimensional structures). In our case, the ultimate goal is to find, for the given new LU u, synsets for which we have the strongest evidence of LUs being in the close hyponymy/hypernymy or near-synonymy relation with u, ideally synsets including its near-synonyms. We assume, though, that the intrinsic errors in data preclude certainty about the exact attachment point for u. Even if synset t appears to fit, we must consider the evidence for t's close surroundings in the hypernymy structure – all synsets located no further than some upper-bound distance d_h. The distance is simply the number of hypernymy links to traverse from t. The evidence for the surroundings is treated as less reliable that for LUs in the central synset t, from the perspective of considering t as the point of attachment. Any information that describes relations of a new lemma with LUs in synsets other than t is related to t only indirectly, by wordnet links. The weight of the context evidence decreases in proportion to the distance.

We consider several sources of heterogeneous evidence for a potential relation of a new lemma with a LU already in the wordnet and thus a relation with some synset. The results of all extraction methods were transformed to sets of LU pairs $\langle x, y \rangle$ such that x and y are semantically related according to the given method and the corpora analysed. There are three groups of sets:

- two sets produced using MSR_{RWF} – the list $MSRlist(y, k)$ of the k units most related to y, and that list restricted to *bidirectional relations*: $MSR_{BiDir}(y, k) = \{y' : y' \in MSRlist(y, k) \wedge y \in MSRlist(y', k)\}$;
- one set generated by the classifier C_H applied to filtering $MSRlist(y, k)$ from LUs not in hypernymy, meronymy or synonymy with x; C_H was trained on the data from plWordNet;
- three sets produced by the manually constructed lexico-syntactic patterns and one set generated by the patterns produced by *Estratto*.

There is only partial overlap among the sources, so we will use them all in expanding the wordnet. We assume that the subsequent methods explore different pieces of partial information available in corpora. We assume, too, that the application of many different methods allows the use of as much lexico-semantic information as possible. Different sources are differently reliable; this can be estimated e.g. by manual evaluation of the accuracy of the extracted pairs. We want to trust the different sources to a different degree: we introduce mechanisms of weighted voting.

The algorithm of Activation-area Attachment

The algorithm is based on the idea of a *semantic fit*: between two lemmas, as representing two LUs linked by a LSR, and between a lemma and a synset, as defining a LU. The fit is identified from all evidence found in corpora. Next, we group synsets which fit the input lemma into activation areas, from which the attachment areas are selected and returned. The attachment areas represent LUs which may have different senses of the given new lemma; the senses are identified from the input data delivered to the algorithm.

Semantic fit between two lemmas. Lemma-to-lemma fit is a function $fit : \mathbf{L} \times \mathbf{L} \rightarrow \{0, 1\}$, where **L** is a set of lemmas, calculated via heuristic voting. Sources of evidence with higher accuracy according to a selective manual evaluation are treated as more reliable and have stronger votes.
$fit(x, y)$ =

- 1 if $x \in C_H(MSRlist(y, k))$ or $x \in MSR_{BiDir}(y, k)$,
- 1 if $\langle y, x\rangle$ or $\langle x, y\rangle$ belongs to at least two sets among $MSR_{RWF}(x, k)$ and sets extracted by patterns,
- 0 otherwise.

The fit *score* is a function $\mathbf{L} \times \mathbf{L} \rightarrow R$.
$score(x, y) =$

- 1 if $fit(x, y)$,
- $MSR_{RWF}(x, y)$ if $x \in MSRlist(y, k)$,
- 0.5 if $\langle x, y\rangle$ or $\langle y, x\rangle$ has been extracted by a pattern with higher accuracy,
- 0.3 if $\langle x, y\rangle$ or $\langle y, x\rangle$ has been extracted by another pattern,
- 0 otherwise.

The weights in the $score$ function have been set experimentally and are based on the manual evaluation of pattern accuracy.

Finding fitting synsets and activation areas. **Phase I** finds all synsets that fit the new lemma. We consider the lemma-to-lemma fit (the new lemma to synset members) and synset contexts. **Phase II** groups the synsets thus found into connected subgraphs – activation areas. For each activation area, the linguist is shown the local maximum of the scoring function; it describes how close the given hit is to the area. The MSR is defined for any lemma pair, but each other source of evidence covers only some pairs. That is why we also introduced a *weak fit* based only on MSR – as opposed to (regular, strong) *fit* which must be based on clues coming from more than one synset member. We expect weak fit to help fill gaps in the description of strong fit between the new lemma and synsets. The missing pieces of more reliable evidence may be due to the limited lexical coverage in the corpora. The weak fit seems to prevent the activation areas from being too fragmented and too small, but it necessarily depends on the predefined threshold $hMSR$ (see below).

We introduce the following notation:

- x is a lemma, representing one or more LUs, to be added to the wordnet, y, y' are lemmas from the wordnet, S, S' – wordnet synsets;
- $hypo(S, n)$, $hyper(S, n)$ are sets of hyponym or hypernym synsets, respectively, of the synset S up to n levels;
- $dist_h(S, S')$ is the number of hypernymy or hyponymy links (depending on the direction) between S and S';
- r is the context radius – it defines the size of the context influencing the calculation of the lemma-to-synset fit (the value was set experimentally to 2);
- $hMSR$ (set experimentally to 0.4) is the threshold defining highly reliable MSR values – it corresponds to the observed high and reliable values of MSR;
- $minMSR$ (set to 0.1) is the MSR value below which associations seem to be based on weak, accidental clues;
- $maxSens$ (set to 5) is the maximal number of presented activation areas (possible attachment areas) – the number of correct proposals is mostly low, so we wanted to keep the number of attachment areas small in order not to clutter the screen.

Phase I. Lemma-to-synset calculation

1. $votes(x, S) = \sum_{y \in S} fit(x, y)$
2. $fit(x, S) =$
   $$\delta_{(h=1)}\left(votes(x, S) + \sum_{S' \in hypo(S,r) \cup hyper(S,r)} \frac{votes(x, S')}{2 * dist(S, S')}\right)$$
   where $\delta : N \times N \to \{0, 1\}$, such that $\delta(n, s) = 1$
   if and only if $(n \geq 1.5 * h$ and $s \leq 2)$ or $(n \geq 2 * h$ and $s > 2)$
3. $$fit(x, S) = \delta_{(h=hMSR)}\Big(\sum_{y \in S} score(x, y) + \sum_{S' \in hypo(S,r) \cup hyper(S,r)} \frac{\sum y \in S' score(x, y)}{2 * dist(S, S')}\Big)$$

Phase II. Identify lemma senses: areas and centres

1. $synAtt(x) = \{\mathbf{S} : \mathbf{S} = \{S : fit(x, S) \vee weak\_fit(x, S)\}$, and $\mathbf{S}$ is a *connected hypernymy graph*$\}$
2. $maxScore(x, \mathbf{S}) = score\,(x, max_{S \in \mathbf{S}} score(x, S))$

3. Remove from $synAtt(y)$ all $\mathbf{S}$ such that $maxScore(x, \mathbf{S}) < minMSR$
4. Return the top $maxSens$ subgraphs from $synAtt(x)$ according to their $maxScore$ values; in each, mark the synset S with the highest $score(x, S)$

The radius r was set to 2 (**Phase I**), because we observed that no extraction method used can distinguish between direct hypernyms and just close hypernyms. The δ function is a mean of non-linear quantisation from the strength of evidence to the decision. We require more *yes* votes for larger synsets, fewer votes for smaller synsets, but always more than one ‘full vote’ must be given – more than one synset member voting *yes*. The parameter h of the δ template relates the function to what is considered to be a ‘full vote’. For weak fit, h is set to the value which signals a very high relatedness for the MSR used.

In **Phase II** we identify continuous areas (connected subgraphs) in the hypernymy graph, those which fit the new lemma x. For each area we find the local maximum of the score function for x. We keep all subgraphs with the synset of the maximum score based on the strong fit (the detail omitted above). From those based on the weak fit, we only keep the subgraphs above some heuristic threshold of the reliable MSR result. We also save for the linguists only a limited number of the best-scoring subgraphs ($maxSens = 5$ – it can be a parameter of the application). We do present all subgraphs with the top synset fit based on the strong fit.

The WordNet Weaver application

The *WordNet Weaver* [WNW] is an expansion of *plWordNetApp* (Piasecki and Koczan, 2007), a wordnet editor developed for the *plWordNet* project and used in its construction (Section 2.4). A separate screen groups most of the added user-perceived functionality – see Fig. 4.8. A linguist sees a list of new lemmas (not yet in the wordnet). A user-selected new lemma u is shown as a green oval. The existing LUs that u fits appear in yellow, orange, red and vivid purple in the increasing order of fitness score. Strong and weak fit is also distinguished by shapes, respectively octagon and rectangle. All fitting synsets together with hypo/hypernymy (arrows point to the hypernyms) links are initially visible to the user: this presents the context of the system's every attachment decision. Only one local maximum per a connected hypernymy subgraph is marked by a blue border. Local maxima – the proposed attachment centres – are graphically linked with the green oval of the new lemma (marked by lines ending with small circles).

The linguist can select any synset present on the screen and then choose a type of lexico-semantic relation, including synonymy, by which it will be associated with the new lemma. A wrong proposal can be rejected, too; in that case, the linguist is asked to select a type and a possible cause of the error. Adding and rejecting removes the circle-

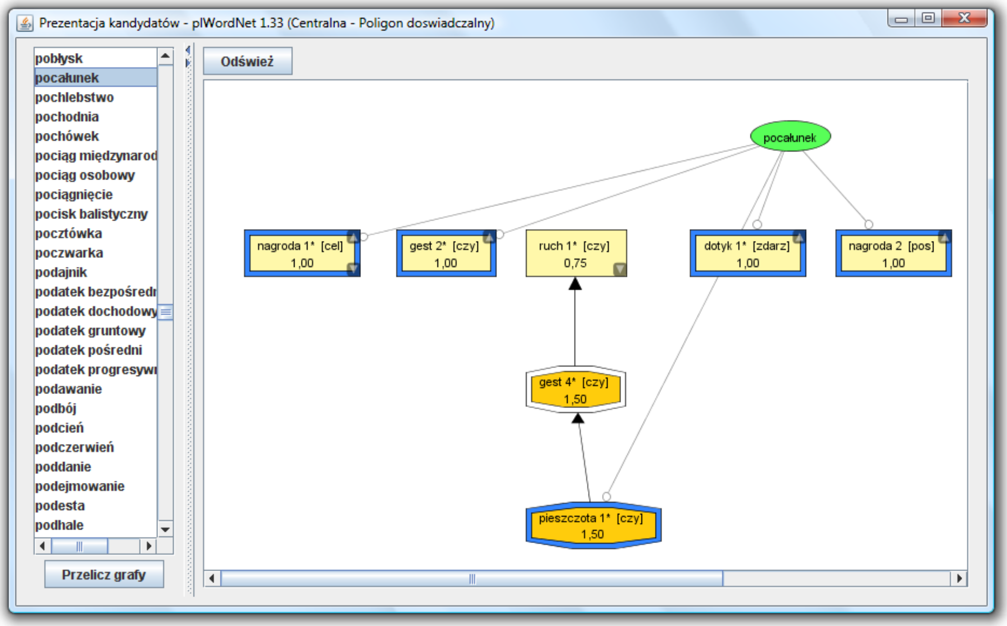


Figure 4.8: WNW's suggestions for 'pocałunek' (*kiss*). Glosses (from left to the right): 'nagroda' *prize/goal*, 'gest' *gesture*, 'ruch' *movement* – the fit, 'dotyk' *touch*, 'nagroda' *prize/possession*, 'pieszczota' *caress*

ending line that links the new lemma oval with the corresponding attachment centre, even in the case of an addition at a high hypernymy distance from the centre. The site of the addition is recorded as the description of the positive proposal. Graphs present on the screen can be selectively unfolded and traversed along hyponymy/hypernymy links (folding/unfolding is accessible via triangle-marked buttons in the top-right corner of a synset symbol), so adding is not limited to synsets marked as fitting the new lemma.

At any moment, the linguist can initiate the Algorithm of Activation-area Attachment [AAA] in order to redefine the attachment areas and centres. Changes affect all new lemmas, but all decisions made so far are kept on the screen. For example, synsets with the new lemma already added are shown as green octagons, together with their relation links.

The total set of new lemmas was automatically divided – by repeatedly running the k-means clustering algorithm from the Cluto package (Karypis, 2002) – into groups that represented mostly quite coherent semantic subdomains. The linguist is shown only one lemma group at a time. She can concentrate on a part of the hypernymy structure. Moreover, such a work procedure is facilitated by the re-computation mechanism, which can improve the attachment proposals by using the information about the new lemmas introduced into the plWordNet structure up to this moment.

The whole attachment screen is embedded in the full plWordNetApp, so the linguist can change any element of the plWordNet database by switching to another panel.

The AAA algorithm runs on the server. On the client side, mainly visualisation is left. WNW is written in Java and can run unchanged on many platforms.

4.5.4 Benefits of weaving the expanded structure

WNW has been designed to facilitate the actual process of wordnet expansion. Its primary evaluation was based on the work of a linguist with rich experience in editing plWordNet, who was adding new nominal lemmas. The candidates came from the same set of 13285 nominal lemmas, which has been defined as a basis for expanding plWordNet during work on MSR extraction, cf Section 3.4.5. The set includes lemmas from a small Polish-English dictionary (Piotrowski and Saloni, 1999), two-word lemmas from a general dictionary of Polish (PWN, 2007) and frequent nouns (>1000) from the joint corpus[14] (≈ 581 million tokens, see Section 3.4.5).

For evaluation purposes, we used 1360 new lemmas divided into subdomains corresponding to *animals* (113 LUs), *food* (170), *people* (323), *people*$_2$ (269), *plants* (81), *places* (243), plus a sample of 161 LUs randomly drawn across all clusters (*rand.* in Table 4.5). Prior to the experiment, the linguist had used only traditional means of her work – electronic dictionaries and corpus browsing. We assumed three types of evaluation:

1. subjective opinions and observations of the linguist collected during actual work over a longer period, 18 person-days,
2. monitoring and analysing the linguist's decisions recorded in the database together with descriptions,
3. automatic evaluation following the general scheme of re-building the existing wordnet by applying the AAA algorithm autonomously.

The linguist's observations

WNW has turned out to be useful in the inclusion of new lemmas given a narrow domain such as `jedz`[15](names of foodstuffs) or `rsl` (plant names). For such lemmas the accuracy was high, and it increased even more as the database grew and as the operation of recomputing the graphs became available. As an example, the program

[14] As described in Section 3.4.5, the joint corpus consists of IPIC (≈ 254 million tokens) (Przepiórkowski, 2004), texts from the electronic edition of a Polish daily *Rzeczpospolita* (≈ 113 million tokens) (Rzeczpospolita, 2008) and a corpus of large Polish texts collected from the Internet (≈ 214 million tokens).

[15] We cite here the original labels assigned to the domains in plWordNet.

suggested fruit and vegetable names (*morela* ‘apricot’ or *pietruszka* ‘parsnip’), names of spices (*tymianek* ‘thyme’) or alcohols (*rum* ‘rum’) as direct or indirect hyponyms of the appropriate existing nodes (*food*, *plant*, *spice*, *alcohol*).

Graph reconstruction did not always work so well. Occasionally, it resulted in suggested link between nouns that were unexpected (*papużka* ‘budgerigar’ → *chomik* ‘hamster’ or *delfin* ‘dolphin’ → *rybka* ‘little fish’) or even random. The system was less helpful for large of general domains such as person or place names – misses were more frequent in such cases[16]. Sometimes LUs were linked quite accurately, but not by hyponymy/hypernymy or synonymy. Instead, the relation was either meronymy (for example, among nouns denoting body parts), fuzzynymy or (less often) relatedness.

Let us note that WNW sometimes also served as a tool for discovering errors in the wordnet. For example, the unit *rój* ‘hive, colony’ was inappropriately linked with the synset *grupa ludzi* ‘a group of people’. The hypernymy tree was missing a node for a group of animals, present in the database but not linked by hyponymy/hypernymy with other LUs from the same semantic field. Similarly, the mislinked LU *holiday* (with hyponyms *Saturday* and *Sunday*) pointed to poorly arranged relations in the synset *day*: *workday* (with hyponyms *Monday* through *Friday*) complements *holiday*. Incomplete hyponym/hypernym trees were also identified. For example, the mislinked new lemma *Koran* ‘The Koran’ showed the absence in plWordNet of the hypernym (*holy scriptures*) as well as co-hyponyms (*The Bible*). A completely haphazard placement of the unit *partykuła* ‘particle’ helped uncover the absence of *part of speech, noun, verb, adjective*, and so on.

Examples

Figures 4.9–4.16 show examples of WNW’s suggestions. The very accurate attachment point (not only the area) suggested for *nikiel* ‘nickel’ exemplifies WNW’s very good performance in such domains like chemical substances and elements, plants or animals. Those are domains of a taxonomical character, but are also well described by the patterns run on the joint corpus. We show in WNW all synsets from a connected subgraph that received a positive score in relation to the subject LU, but only the local maximum (a synset marked by the blue border) is the final singular attachment point. The whole subgraph represents the attachment area.

[16] This subjective observation comes from the linguist who worked with WNW. It has been partially contradicted by the statistical data recorded in the database that registered the linguist’s every decision. The data show that the most hits at the level of new lemmas was observed in the domain of persons, so the observation may have been caused by the lower number of direct hits in this domain. This example is a good illustration of possible discrepancies between objective statistically-based evaluation and the usability of a tool for the users.

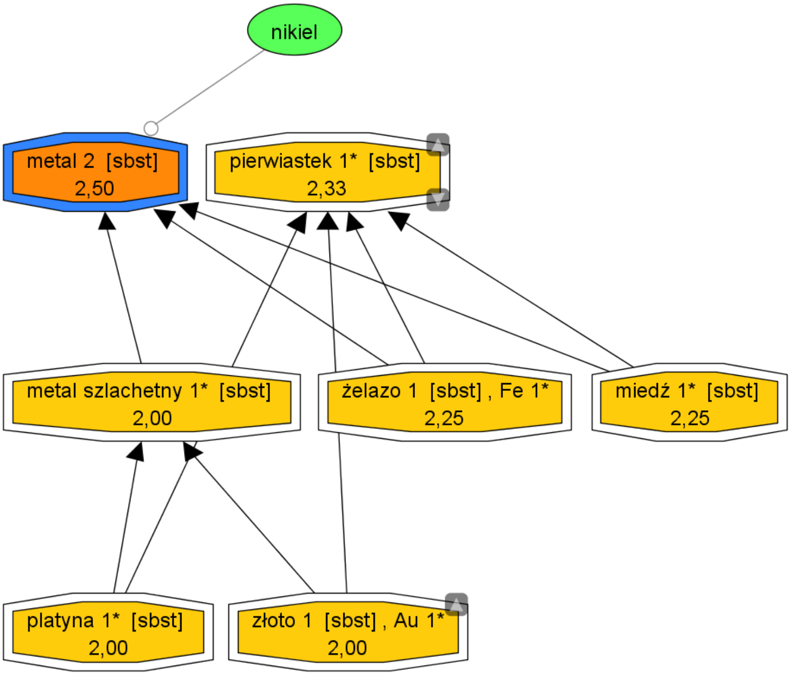


Figure 4.9: WNW's very accurate suggestions for 'nikiel' (*nickel*). Glosses: *metal* 'metal', *pierwiastek* 'element', *metal szlachetny* 'precious metal', *żelazo* 'iron', *miedź* 'copper', *platyna* 'platinum', *złoto* 'gold'

For the LU *semestr* 'semester' (Figure 4.10) the attachment has been placed one level lower than the correct hypernym. The high score value of *miesiąc* 'month' may have been increased by the context in which we can find the three months comprising a winter semester[17] at Polish universities. The grey octagon of *czas* 'time' is a higher-level hypernym, not considered by AAA but unfolded manually to show the context.

WNW proposed three sense for *statuetka* 'figurine', Figure 4.11. The most accurate sense seems to be a hyponym of *rzeźba* 'sculpture' but the popularity of the Oscar ceremony, makes *statuetka* 'figurine' acceptable as a hyponym of *nagroda* 'award'. Only the third sense proposed is incorrect but it is semantically related to the first one[18].

The way in which WNW has arrived at the proposal of the attachment point for *karafka* 'carafe' can be almost followed on the diagram presented in Figure 4.12. The

[17] An academic year in Poland is divided into a winter semester and a summer semester.

[18] In the present version of plWordNet *rzeźba* 'sculpture' and *nagroda* 'award' are not connected by a hypernymic path. The top hypernym for *rzeźba* is *dzieło* 'work' while for *nagroda* 'award' it is *rzecz, przedmiot* 'thing, object'.

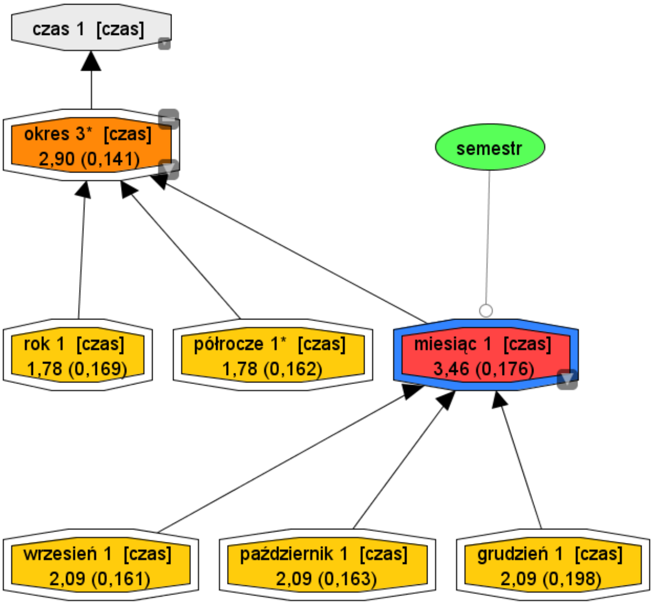


Figure 4.10: An almost correct attachment suggested by WNW for *semestr* ‘semester’. Glosses: *czas* ‘time’, *okres* ‘period’, *rok* ‘year’, *półrocze* ‘semester, half-year’, *miesiąc* ‘month’, *wrzesień* ‘September’, *październik* ‘October’, *grudzień* ‘December’

considered LU received high scores in relation to many glass dishes, but those were outweighed by synsets in the lower part that denote different types of bottles.

The first attachment proposed for *dzięcioł* ‘woodpecker’ is predictable (Figure 4.13), since WNW’s precision is high for animal names. The second proposed sense may seem to be an error but *sikorka* ‘tit’ is not so far from the meaning of *dzięcioł* ‘woodpecker’ in the school jargon (someone who is believed to spend much time memorising school material). So, the second sense *is* a person, thus a distant hyponym of *sikorka* ‘girl’. Still, what matters is probably only the accidental similarity of *dzięcioł* and *sikorka* as birds.

Attachments for *urna* ‘urn’ presented in Figure 4.14 have been generated only based on weak fitness (MSR$_{RWF}$) values. On average, attachment proposals computed from weak fitness are of lower quality, but not in this case. Two proposed senses show a case of regular polysemy characteristic for many Polish LUs in plWordNet that denote a container. They have a thing sense and a place sense.

Two of the four attachment points (all based on weak fitness) for *sztaba* ‘bar (of metal)’ are the result of an error introduced by the TaKIPI tagger when preprocessing

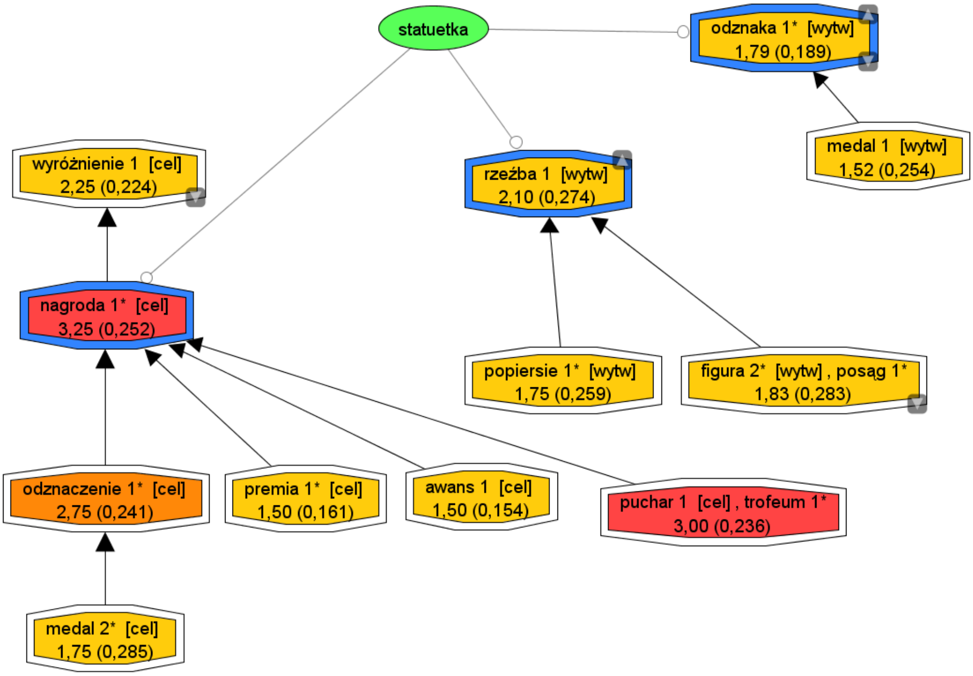


Figure 4.11: Two correct senses and one closely related, which WNW predicted for *statuetka* 'figurine'. Glosses: *wyróżnienie* 'distinction', *nagroda* 'award', *odznaczenie* 'honour, decoration', *premia* 'bonus', *puchar* 'cup', *medal* 'medal', *rzeźba* 'sculpture', *popiersie* 'bust', *figura* 'figure = image', *odznaka* 'honour, decoration'

the joint corpus[19]. Several word forms of the LU *sztaba* are shared with the word forms of the LU *sztab* 'staff = command centre':

- selected word forms of *sztaba*:
  $sztaby_{nmb=sg,case=gen}$, $sztaby_{nmb=pl,case=nom}$, $sztabie_{nmb=sg,case=dat}$, $sztabie_{nmb=sg,case=loc}$, $sztabie_{nmb=sg,case=loc}$, $sztabie_{nmb=pl,case=voc}$, $\ldots$;
- selected word forms of *sztab*:
  $sztaby_{nmb=pl,case=nom}$, $sztaby_{nmb=pl,case=voc}$, $sztabie_{nmb=sg,case=loc}$, $sztabie_{nmb=sg,case=voc}$.

The other two attachment proposal for *sztaba* are semantically related but not very accurate. Both are based only on weak fitness. That is why we differentiate strong and weak fitness by different symbols on the screen. This example also shows the role of multiple criteria in WNW.

[19] Such errors are rather rare, good news given TaKIPI's solid but not stellar accuracy.

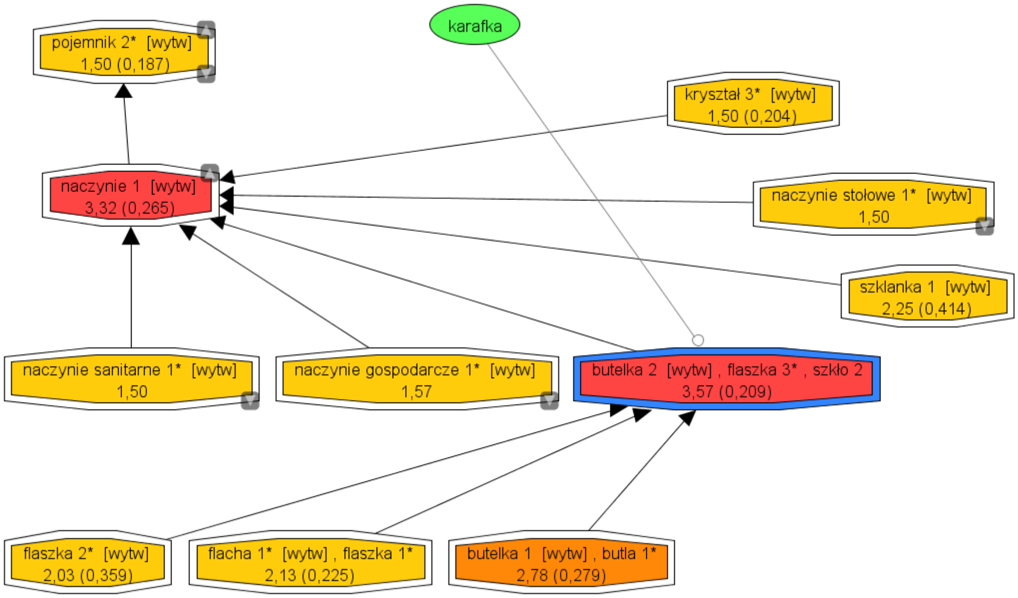


Figure 4.12: A correct WNW decision made for *karafka* 'carafe' on the evidence collected from the context. Glosses: *pojemnik* 'container', *naczynie* 'pot = utensil', *naczynie sanitarne* 'sanitary pot', *naczynie gospodarcze* 'household container', *butelka, flaszka, szkło* 'bottle, bottle, (glass) bottle', *butelka, butla* 'baby bottle', *kryształ* 'crystal (container)', *naczynie stołowe* 'tableware', *szklanka* 'glass = container'

The last example (Figure 4.16) illustrates the problems that the present version of WNW has with Polish gerunds. Their specific syntactic behaviour prevents the MSR extraction method and the patterns from producing good results. We plan to work on a dedicated solution. Figure 4.16 is also an honest admission that WNW is sometimes not helpful for linguists, but at any moment they can switch to completely manual editing of the plWordNet database for any new lemma.

In the following points we investigate the statistical measures of WNW's accuracy.

Analysis of the linguist's work

WNW stores information on the linguist's every decision. For positive decisions, the application records the distance along hyponymy/hypernymy links between the point of attachment and the proposed attachment centre, plus a possible linguist's comment. For negative decisions, the linguist is asked for the error type and its possible cause. Seven error types include such situations as a tagging error or an unlisted sense that the linguist has to add manually.

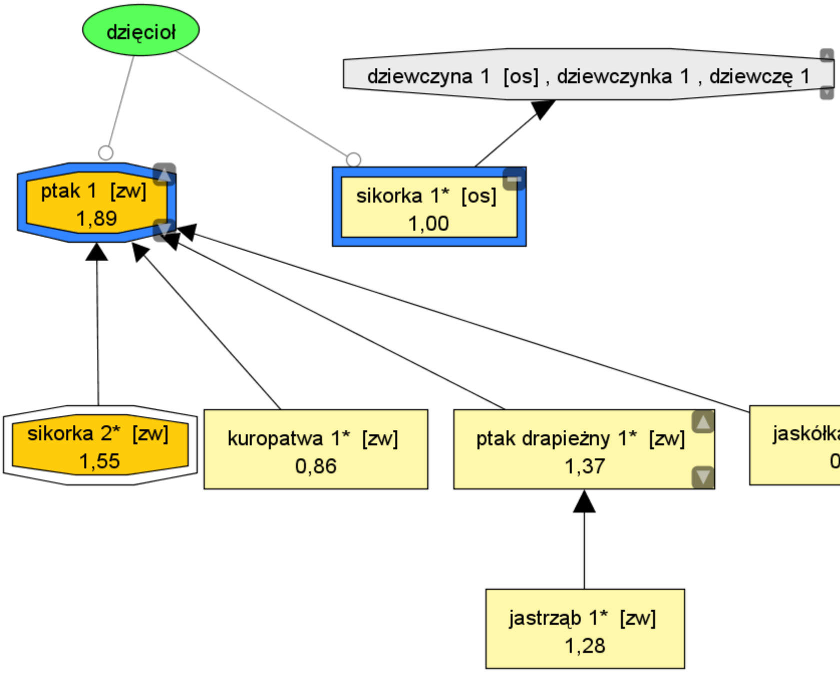


Figure 4.13: WNW's accurate suggestion, and one indirectly related to a informal sense, for *dzięcioł* 'woodpecker'; the informal sense is 'one who cons (habitually learns by heart)'. Glosses: *ptak* 'bird', *kuropatwa* 'partridge', *ptak drapieżny* 'carnivore bird', *jaskółka* 'swallow', *jastrząb* 'hawk', *sikorka* 'tit, chikadee; informally: adolescent girl', *dziewczyna* 'girl'

Attachment errors were mostly caused by text preprocessing or by the AAA algorithm itself. Since plWordNet is still very much under development, a wrong attachment can be caused by a flaw in the wordnet structure. In order to distinguish among several different causes, we asked the linguist to try to identify the possible cause for every wrong attachment. Eight cause types include a missing hypernym or substructure, a duplicated synset or a wrong synset element.

Precision and recall calculated for the recorded decisions appear in Table 4.5. Precision was calculated in relation to

- different acceptable distances: P_1 – exact attachment to the local maximum synset (by synonymy or hyponymy) versus P_H – at any distance accessible by links in a given subgraph,
- acceptable types of links: P_{H+M} – meronymic links are counted as positive too, because they also are in the linguist's focus,

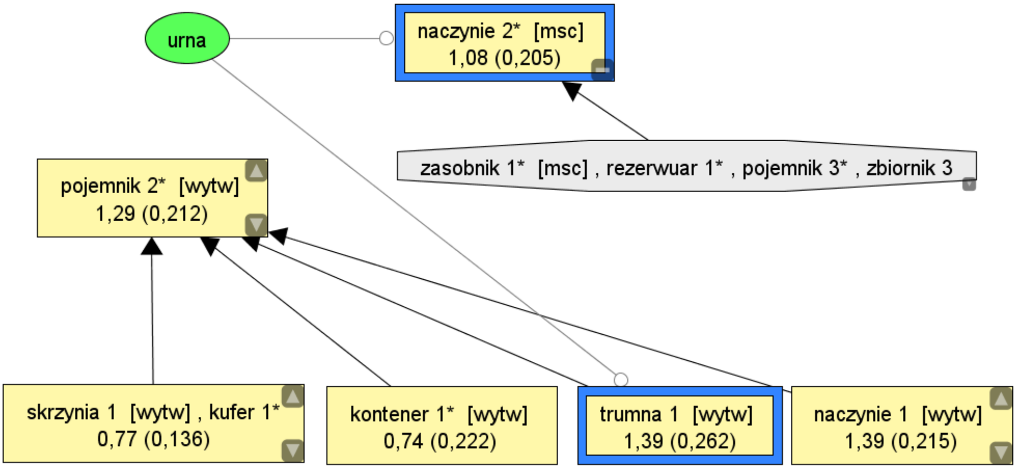


Figure 4.14: Two close attachments with regular polysemy that WNW suggested for *urna* 'urn' based only on weak fitness. Glosses: *pojemnik* 'container', *skrzynia* 'chest', *kontener* 'container', *trumna* 'coffin', *naczynie* 'container', *zasobnik, rezerwuar, pojemnik, zbiornik* 'container/tank, reservoire, container, container/reservoir/receptacle'

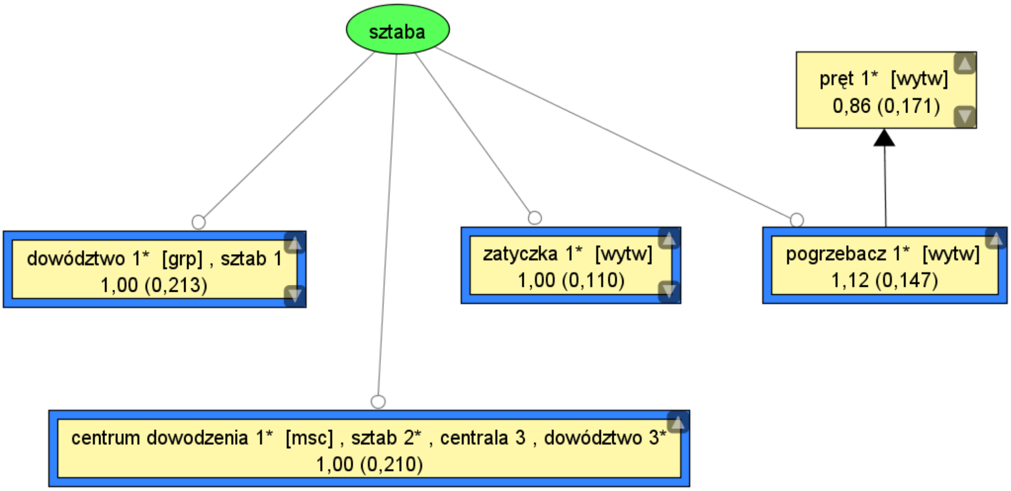


Figure 4.15: Two wrong attachments for *sztaba* 'bar' caused by the TaKIPI tagger error in identifying LUs (*sztaba* and *sztab* 'staff') with overlapping paradigms. Glosses: *dowództwo* 'command', *centrum dowodzenia* 'command centre', *centrala* 'centre, central', *zatyczka* 'plug', *pogrzebacz* 'poker', *pręt* 'rod'

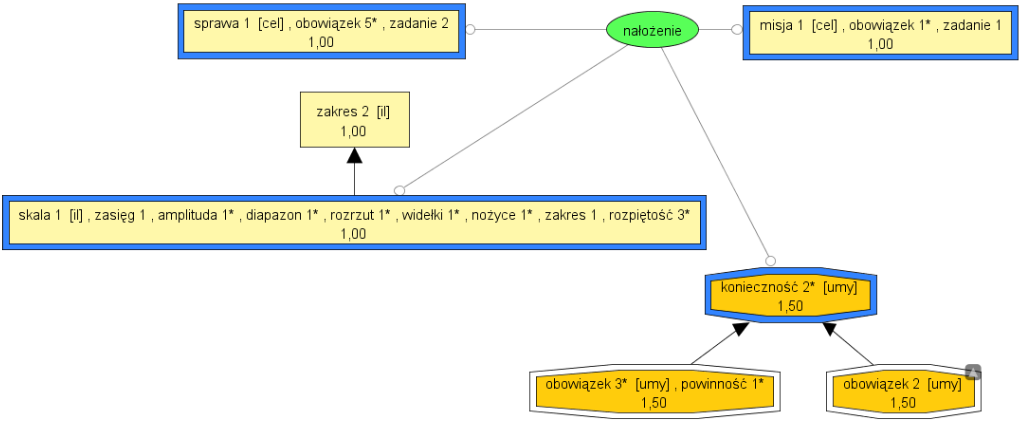

sprawa 1 [cel] , obowiązek 5* , zadanie 2
1,00
nałożenie
misja 1 [cel] , obowiązek 1* , zadanie 1
1,00
zakres 2 [il]
1,00
skala 1 [il] , zasięg 1 , amplituda 1* , diapazon 1* , rozrzut 1* , widełki 1* , nożyce 1* , zakres 1 , rozpiętość 3*
1,00
konieczność 2* [umy]
1,50
obowiązek 3* [umy] , powinność 1*
1,50
obowiązek 2 [umy]
1,50

Figure 4.16: Completely wrong attachments generated for the gerund *nałożenie* 'imposition (an act of imposing)' by WNW. Glosses: *sprawa, obowiązek, zadanie* 'cause, duty, task', *misja* 'mission', *skala, zasieęg, amplituda, . . .* 'scale, range, amplitude . . . ', *zakres* 'range', *konieczność* 'necessity', *obowiązek, powinność* 'duty, obligation'

- different measure of success: $P_{\geq 1}$ is the percentage of new lemmas for which at least one suggestion was successful as $H + M$.

Recall was calculated thus:

$$\frac{accepted\ attachments}{accepted\ attachments\ +\ senses\ added\ by\ the\ linguist}$$

group	P_1	P_H	P_{H+M}	$P_{\geq 1}$	R
all	14.86	34.58	36.35	80.36	75.24
rand.	13.10	27.62	30.48	59.12	55.77
animals	18.61	45.89	48.05	91.74	86.18
food	19.52	34.76	38.57	65.68	61.09
people	14.04	42.11	43.05	86.02	79.39
people$_2$	14.16	38.67	39.80	95.38	90.40
plants	20.28	38.71	43.78	83.95	80.77
places	16.64	32.51	36.52	74.58	70.57
diseases	11.38	31.14	34.13	77.27	67.53

Table 4.5: Precision (P) and recall (R) [%] of attachment measured during linguist's work

The precision and recall are much higher for such coherent subdomains such as plants or animals than for the randomly drawn sample. In the former case, the linguist often accepts the suggestion immediately. Gerunds were a problem spot, with a decreased accuracy in both MSR_{RWF} and patterns. The exact precision values may seem low but the number of possible attachments is set to a rather high 5 – to show the linguist more extracted senses – so it impairs precision. The measure $P_{\geq 1}$ also shows that at least one proper attachment area was identified in the majority of LUs. Even for the worst random sample, proposals for 59.12% of new lemmas were found worth examining, as they include helpful suggestions. The numbers do not show how the tool can inspire the user, draw her attention to less obvious or domain-dependent senses, reveal peculiarities in the wordnet state and so on.

L	All_S	All_S	All_{S+W}	One_S	One_W	One_{S+W}	$Best_{P\geq 1}$
0	26.65	7.90	16.46	45.80	16.24	34.96	42.81
1	35.76	14.50	24.21	58.73	28.96	47.81	61.19
2	42.87	21.39	31.20	67.69	40.51	57.72	75.02
3	48.31	27.36	36.93	73.58	51.08	65.33	81.96
5	53.52	34.78	43.34	78.46	58.51	71.14	86.18
6	57.59	43.59	49.99	81.52	64.58	75.31	89.82
7	61.09	49.90	55.01	83.56	70.45	78.75	91.16
8	63.38	53.71	58.13	84.47	73.58	80.47	92.49
9	65.27	56.55	60.53	85.03	75.73	81.62	93.12
10	66.07	58.86	62.16	85.26	78.28	82.70	93.54

Table 4.6: The accuracy [%] of plWordNet reconstruction; L – the distance from the original synset, *S* and *W* mean strong and weak fitness, respectively

WordNet reconstruction

In the automatic evaluation, we wanted to check the ability of the AAA algorithm to reconstruct parts of plWordNet. The method is meant to expand the existing core structure of a wordnet, so we identified 1527 LUs in the lower parts of the hypernymy structure as a basis for the evaluation. In order to introduce as little bias as possible, 10 LUs were removed from the plWordNet structure in one step of the evaluation. The C_H classifier component was trained without the removed LUs and the AAA algorithm was run to attach the processed LUs.

There are many synsets in plWordNet with a single LU. This makes the evaluation of LUs in such synsets problematic. If we removed singleton synsets, we would artificially – and dramatically – alter the overall structure of plWordNet and so introduce an unwanted bias. That is why we decided to remove only the LUs and to leave empty synsets in the modified plWordNet.

We assumed three strategies for evaluating the AAA algorithm's proposals:

- *All* – all proposals are evaluated;

- *One* – only single highest-scoring attachment site is evaluated; this strategy was introduced mainly for comparison with other approaches (but it is unnatural from the point of view of the linguists' work);

- $Best_{P\geq 1}$ – one closest attachment site is evaluated (similarly to the $P_{\geq 1}$ in Sec. 4.5.4).

Table 4.6 presents the result with a distinction between strong (marked S) and weak fitness (W). As expected, the accuracy of suggestions based on strong fitness is significantly higher then for weak fitness. Because of the intended use, we assumed that not only direct hits are useful – if the proposal is close enough to the correct place in plWordNet structure, then it is also a valuable suggestion. The same applies if there is meronymy or holonymy between the suggested and correct synset.

The results are encouraging. Almost half of the suggestions based on strong fitness are in the close proximity of the correct place in wordnet structure. If making only one suggestion was required, the accuracy was boosted to 73.58%. For our goal, this is an artificial constraint, but it shows how well the AAA algorithm would behave in a fully unsupervised way. Our ultimate goal, though, is to create a tool for supporting a linguist's work, so the result for $Best_{P\geq 1}$ strategy shows more meaningful data: for how many words there is at least one useful suggestion. The AAA algorithms suggested at least one strictly correct attachment site for 42.81% words, or for 81.96% words if we consider that close proposals are also useful.

Comparison to other ways of automatic expanding a wordnet can be misleading. That is because our primary goal was to construct a tool that facilitates and streamlines the linguists' work. Still, even if we compare our automatic evaluation with the results in (Widdows, 2003) during comparable tests on the PWN, our results seem to be better. For example, we had 34.96% for the highest-scored proposal (One_{S+W} in Table 4.6), while Widdows reports a 15% best accuracy for a "correct classifications in the top 4 places" (among the top 4 highest proposals). Our similar result for the top 5 proposals is even higher, 42.81%. The best results reported by Alfonseca and Manandhar (2002) and Witschel (2005) are also at the level of about 15%, but were achieved in tests on a much smaller scale. Witschel also performed tests only in two selected domains. The algorithm of Snow et al. (2006), contrary to ours, can be applied only to probabilistic evidence.

We made two assumptions: attachment based on the activation area and the simultaneous use of multiple knowledge sources. The assumption appears to have been successful in boosting the accuracy above the level of the MSR-only decisions (which is roughly represented in our approach by weak fitness).

WNW seems to improve the linguist's efficiency a lot, but longer observations are necessary for a reliable justification.

The AAA algorithm is overburdened with parameters. Further research is required to find either a simplified form or an effective method of parameters optimization.

Chapter 5

Polish WordNet Today and Tomorrow

5.1 Weaving the Full-fledged Structure

The two preceding chapters present a suite of semantic extraction tools and the WordNet Weaver [WNW] application which makes them available. We have developed this toolkit to support linguists' work on expanding the core plWordNet into a full-fledged structure of Polish WordNet [plWordNet] version 1.0. The expansion process (Sections 4.5.3 and 4.5.4) allowed us to use WNW, and the tools it offers, on a realistic scale – and to evaluate it in practice.

Section 2.5 presents the core plWordNet. We started the expansion process with 13285 one- and two-word nominal lemmas, acquired mainly from dictionaries and supplemented by the most frequent lemmas in the joint corpus of 581 million tokens (Section 3.4.5 describes the corpus and the set); that included 7200 lemmas not in the core plWordNet. Preliminary experiments with WNW showed that some lemmas taken from the dictionaries has inadequate support in the corpus, so in the end only about 5500 new lemmas were effectively used for plWordNet expansion. That is because our primary dictionary source (Piotrowski and Saloni, 1999) is small but contains a number of rare lemmas; two-word lemmas – acquired from (PWN, 2007) – are inevitably less frequent than one-word lemmas; and the corpus is not sufficiently balanced. The extraction tools also introduced a bias:

- lexico-morphosyntactic constraints in the MSR construction have been boosted for precision, so some association occurrences have been left unrecognised;
- lexico-morphosyntactic patterns can recognise a limited variety of lexico-syntactic structures, and in any event the analysed lemma pairs can only be harvested if they occur in one of the recognised structure types.

All this made it necessary to supplement the lemma list with lemmas occurring at least 500 times in the joint corpus, if we were to achieve the intended size of the wordnet. The completed list had 9011 new nominal lemmas, 15096 in total[1].

[1] This included lemmas from the core plWordNet required by the extraction tools, in particular in the extraction of relation instances by patterns.

The expansion process was therefore slightly biased towards a data driven approach[2]. Such a bias is not necessarily a drawback, however, because – assuming moderate influence of the linguists' decisions – the resulting resource represents a structure of lexico-semantic relations implicitly present in the language data.

The operation of pattern-based methods was also limited to the list of 15096 selected lemmas – its late version supplemented by the frequent lemmas from the joint corpus. (Support for this design decision comes from preliminary experiments, in which we applied pattern-based methods to an unrestricted, full list of lemmas collected from the joint corpus. In the absence of any restrictions, there were very many extracted instances, at the cost of much lower precision.)

An improvement in the number of dictionary lemmas covered can naturally be had if there is a larger and better balanced corpus to cover more domains and more examples per word sense. This is not a realistic wish, but there still is much room in semantic extraction for increased use of what *is* present in texts. For example, MSR construction could easily benefit from a deeper analysis of the lexico-syntactic structure afforded by a syntactic analyser.

We only applied the semi-automatic process to nominal lemmas. The present version WNW strongly depends on the hypernymy structure, rich for nouns, but quite limited for verbs and very rare for adjectives. There simply was no basis for the identification of attachment areas. (In the future versions of WNW, we plan to use more relation types in generating suggestions.) Also, few lexico-syntactic patterns work for verbs. It is a serious open problem to find effective use of information coming from single occurrences of verbal and adjectival lemmas in a style of pattern-based approaches.

Here is how we have organised around WNW the process of expanding plWordNet.

1. We collected and morpho-syntactically preprocessed a large corpus (Section 3.4.3).

2. From the corpus, we extracted data sets describing lexico-semantic relations; we applied all constructed and experimentally verified automated extraction tools: the Measure of Semantic Relatedness MSR_{GRWF}, manual patterns and Estratto (Section 4.3).

3. A classifier for lexico-semantic relation (Section 4.5.1) was trained on the then-current state of the wordnet and the data sets from the previous step.

4. New lemmas were clustered using MSR_{GRWF} and an off-the-shelf clustering tool; we applied the Cluto package (Karypis, 2002).

[2] It is important to note that the ultimate decisions belonged to linguists, who could also freely add relation instances not suggested by WNW.

5. Selected groups of new lemmas were loaded into WNW and the Algorithm of Activation-area Attachment [AAA] was run to generate suggestions of attachment areas.

6. Linguists worked freely with the lemma groups; they browsed suggestions in any order and edited the wordnet structure.

7. At any moment of the process, linguists could re-run AAA to get perhaps better suggestions for those new lemmas that have not been edited yet.

8. Linguists notified the coordinator about finishing work with particular groups; the coordinator then could analyse the results using the same WNW system (accessing it via the Internet, just like the linguists).

The whole process of extracting data sets – sources of evidence for AAA – performed in steps 1-2 took approximately 25 days on a standard PC (3GHz, 4GB RAM, one single-core processor). The time could be reduced to 2-4 days by applying a grid of at least several PCs. This *one-time* operation is computationally very intensive, but it prepares all data sets except classifiers at the beginning of a long-term expansion process. This is done once per each list of new lemmas, independent of the size of the list. Classifier training, to be repeated several times with the increasing size of the wordnet, it is much less computationally demanding than the other tasks. AAA is performed on the server, not on the linguists' PCs. It takes 10-20 minutes on a PC-class server.

Clustering (step 4) is optional from the point of view of the WNW application, which can work efficiently with a list of several thousand new lemmas. Clustering is necessary for people: a huge flat list is just too difficult to comprehend, and it is practically impossible to organise around it work lasting several weeks.

The idea behind clustering was to divide the initial list into lemma groups in such a way that each group consists of lemmas with senses belonging to one domain common to all of them (at least the intersection of the lemma senses should belong to one domain). There is no perfect clustering algorithm, but manual grouping would be too labourious to be feasible. We applied an off-the-shelf implementation of clustering algorithms in the Cluto package (Karypis, 2002). The input to the clustering algorithms were values which describe semantic relatedness of lemma pairs acquired from MSR$_{GRWF}$. We experimented with different algorithms. After a manual inspection of the results, we selected *graph-based clustering*. We did not evaluate the quality of clustering exhaustively: the mechanism played only a minor, supporting role. Due to the properties of the clustering algorithms, we repeated the process several times, each time getting some groups and a large set of 'outliers', which was next the input to another run. The obtained groups were loaded into WNW – all in all, 92 groups were constructed.

akacja ‘black locust (false acacia)’, *bez* ‘lilac’, *bluszcz* ‘ivy’, *brzoza* ‘birch’, *buk* ‘beech’, *busz* ‘bush’, *bylina* ‘perennial’, *cedr* ‘cedar’, *choinka* ‘Christmas tree’, *chrust* ‘dry twigs’, *chryzantema* ‘chrysantemum’, *chwast* ‘weed’, *cis* ‘yew’, *cyprys* ‘cypress’, *darnia* [a lemmatisation error; should be *darń* ‘sward’], *drzewko* ‘(small) tree’, *drzewostan* ‘forestation’, *fiołek* ‘violet’, *gałązka* ‘twig’, *gęstwina* ‘thicket’, *girlanda* ‘garland’, *głóg* ‘hawthorn’, *goździk* ‘carnation’, *hiacynt* ‘hyacinth’, *irys* ‘iris’, *jabłoń* ‘apple tree’, *jawor* ‘sycamore maple’, *jemioła* ‘mistletoe’, *jeżyna* ‘blackberry’, *jodła* ‘fir’, *kaktus* ‘cactus’, *klon* ‘maple’, *koniczyna* ‘clover’, *konwalia* ‘lily of the valley’, *kora* ‘bark’, *korzenie* ‘roots’, *krokus* ‘crocus’, *kwiatek* ‘(small) flower’, *leszczyna* ‘hazel’, *lilia* ‘lily’, *listowie* ‘foliage’, *łyko* ‘phloem’, *mech* ‘moss’, *modrzew* ‘larch’, *narcyz* ‘narcissus’, *orchidea, oset* ‘orchid, thistle’, *osika* ‘aspen’, *palma* ‘palm tree’, *papirus* ‘papyrus’, *paproć* ‘fern’, *platan* ‘plane tree’, *pnącz* [a lemmatisation error; should be *pnącze* ‘creeper’], *pnącze* ‘creeper’, *pokrzywa* ‘nettle’, *polano* ‘log’, *rododendron* ‘rhododenron’, *roślinność* ‘vegetation’, *sadzonka* ‘seedling’, *sitowie* ‘rush’, *słonecznik* ‘sunflower’, *sosna* ‘pine’, *stokrotka* ‘daisy’, *szałwia* ‘sage’, *szyszka* ‘cone’, *ściernisko* ‘stubble field’, *świerk* ‘spruce’, *topola* ‘polar’, *trzcina* ‘reed’, *tulipan* ‘tulip’, *wiąz* ‘elm’, *wić* ‘runner’, *wieniec* ‘wreath’, *wierzba* ‘willow’, *winorośl* ‘grape vine’, *wodorost* ‘alga, seaweed’, *wrzos* ‘heather’, *zarośle* ‘thicket’, *źdźbło* ‘blade (of grass)’, *żonkil* ‘daffodil’, *żywopłot* ‘hedge’

aktówka ‘briefcase’, *atrament* ‘ink’, *bagaż* ‘luggage’, *bibuła* ‘blotting paper’, *bibułka* ‘tissue paper’, *bloczek* ‘notepad’, *cerata* ‘oilcloth’, *chlebak* ‘haversack’, *cyrkiel* ‘compass (for drawing)’, *długopis* ‘ball-point pen’, *dzianina* ‘hosiery’, *filc* ‘felt’, *grzechotka* ‘rattle’, *gumka* ‘eraser’, *hamak* ‘hammock’, *juk* ‘saddle bag’, *kabura* ‘holster’, *karton* ‘carton’, *klocek* ‘(toy) block’, *kojec* ‘pen (for a child)’, *kołyska* ‘cradle’, *koperta* ‘envelope’, *kredka* ‘crayon’, *leżak* ‘deck chair’, *łóżeczko* ‘(small) bed’, *markiza* ‘awning’, *mat* ‘mate, matte’, *mata* ‘mat’, *muślin* ‘muslin’, *namiot* ‘tent’, *nosze* ‘stretchers’, *nożyczki* ‘scissors’, *ołówek* ‘pencil’, *otomana* ‘sofa’, *paczuszka* ‘(small) package’, *pakunek* ‘package’, *pergamin* ‘parchment’, *perkal* ‘gingham’, *pędzel* ‘brush’, *pierzyna* ‘duvet’, *plastelina* ‘plasticine’, *poduszeczka* ‘(small) pillow’, *przybór* ‘implement’, *saszetka* ‘sachet’, *segregator* ‘binder’, *siodełko* ‘seat’, *skakanka* ‘skip rope’, *skoroszyt* ‘folder’, *spinacz* ‘paper clip’, *stalówka* ‘nib’, *stołek* ‘stool’, *szala* ‘tray (in scales)’, *sztaluga* ‘easel’, *tłumok* ‘(large) bundle’, *tobół* ‘(large) bundle’, *tornister* ‘knapsack’, *tusz* ‘ink’, *włóczka* ‘yarn’, *woreczek* ‘(small) sack’, *worek* ‘sack’, *wór* ‘(large) sack’, *wyściółka* ‘lining, padding’, *zawiniątko* ‘bundle’, *zwitek* ‘scroll, wad, roll’, *zwitka* [a lemmatisation error; should be *zwitek* ‘scroll, wad, roll’]

Figure 5.1: Examples of groups of new lemmas created by automatic clustering

It was very hard to find a pure one-domain group, but most groups seem to fall into only two-three domains. Figure 5.1 shows two examples. This had positive influence on the expansion process. Skimming a group usually sufficed to identify its main domains, so we could direct the expansion process first toward the missing parts in the hypernymy structure. The linguists could concentrate on a few domains and gradually expand the given hypernymy subgraphs while working with a given group. After adding some LUs to the given domain, AAA could be rerun to recompute suggestions for the still unedited lemmas; in narrow domains with deeper hypernymy structure, such as *food* or *clothing*, this increased the accuracy of suggestions and facilitated the linguists' work. Later on, experienced linguists were able to decide for which group the slightly time-consuming recomputation is worth doing.

WNW was designed as a plug-in to the wordnet editor (Section 2.4). AAA-generated suggestions (step 6) presented as shown in Section 4.5.4 appear in a panel, so the linguist can switch to the editor in any moment, and can change the wordnet structure as needed.

In order to increase the quality of the results, we assumed expansion in two stages: editing the suggestions followed by a verification of the results by the coordinator. The editing was local in a usually small part of the hypernymy structure. The new LUs were woven into the structure by hypernymy links. Linguists often adjusted several levels of the hypernymy hierarchy, but again locally, without concern for the overall shape of the plWordNet structure. The coordinator brought in a broader perspective, trying to tie up and merge new hypernymy branches emerging from local expansion work. The existing structure was also corrected in many points, as new synsets were appearing.

As a result of the expansion process, the core plWordNet grew by 8316 new lemmas, 10537 new LUs, 8729 synsets and 11063 instances of lexico-semantic relations; this took 3.4 person-months. There were many improvements to the core plWordNet structure. While we estimate that the expansion process was sped up 5–6 times in comparison to purely manual work, no proper comparison with manual expansion was performed (for example, we could not afford working with the same lemma list independently in two different ways). Longer observations and further experiments are necessary for a reliable justification.

The precision of AAA-generated suggestions when measured using the complete expansion results was lower than reported for the test part in Table 4.5, page 162. Here are the percentages of new lemmas for which at least one suggestion (denoted $P_{\geq 1}$ on page 160) was successful[3]:

- 63.76% – all nominal suggestions,
- 61.43% – suggestions generated for new lemmas described as gerunds or ambiguous between gerunds and nouns,
- 64.12% – new lemmas which the morphological analyser *Morfeusz* (Woliński, 2006) unambiguously recognised as nouns.

The results are significantly lower than 80.36% which Table 4.5 reported for all lemmas in the test set. We attribute the drop in accuracy mainly to less frequent new lemmas included in groups processed at the end of the expansion process; the automatic description of such lemmas tends more often to be inadequate. AAA also gives lower results for gerunds, but the difference between the $P_{\geq 1}$ accuracy for gerundial lemmas and substantive lemmas is not as significant as it appears from manual inspection. The reason is that there are on average more suggestions generated for a gerundial lemma

[3] With hypernymy and meronymy links counted as positive – see page 160, the description of the symbol $H + M$.

than for substantive lemmas. Nevertheless, valuable suggestions – of use to linguists – have been generated most of the new lemmas, including even those relatively rare.

5.2 plWordNet at Three

The present state of plWordNet – version 1.0 – is the effect of the productive semi-automatic expansion discussed in Section 4.5.3. We note that it was the linguists who have introduced all new synsets and instances of lexico-semantic relations, following automatically generated suggestions.

As discussed in Section 2.5, we chose to measure the size of plWordNet in lemmas and lexical units, but Table 5.1[4] shows synset numbers too. This facilitates the comparison with other wordnet descriptions in the literature.

	Nouns	Verbs	Adjectives	*All*
Lemmas				
All	14131	3497	2636	20223
Monosemous	10839	2777	1924	15477
Polysemous	3292	720	712	4746
LUs	18611	4498	3881	26990
Synsets	13675	1860	2160	17695

Table 5.1: The size of plWordNet, version 1.0

Adverbial LUs have not been included in the first version of plWordNet. Instead, we increased the number of nominal and verbal LUs, with the strong emphasis on the former: there are 1.54 times more nominal lemmas than verbal and adjectival lemmas together. The corresponding ratio in PWN is still much higher ("WordNet 3.0 database statistics" in (Miller et al., 2007)): there are 3.57 times more nominal "strings" than verbal and adjectival ones together. The data collected from the joint corpus (Section 4.5.4), automatically processed by a morphosyntactic tagger, show the ratio: 1.45 nominal lemma (including several hundred multiword LUs from the list prepared for expanding plWordNet, cf Section 4.5.4) per one verbal or adjectival lemma. There is only a moderate nominal LU bias in plWordNet, compared to the state in PWN and the corpus.

There are more LUs in plWordNet than synsets. That is because one LU belongs to exactly one synset but a synset can group several LUs (Table 5.1). Reporting the distinction between the monosemous and polysemous lemmas follows the practice of PWN (Miller et al., 2007).

[4] The counts describe the state of plWordNet at the moment of writing the book. See plwordnet.pwr.wroc.pl for the up-to-date numbers.

The total number of LUs described in plWordNet 1.0 is a little above the range declared in the project proposal (15–25 thousand LUs). It compares quite favourably with the wordnets created during the second phase of the EuroWordNet [EWN] project (Vossen et al., 1999, p. 7). We selected the older versions of the corresponding wordnets: they were created from scratch (like plWordNet), and the EWN 2.0 project lasted several years (like plWordNet). The comparison must be taken with a grain of salt, because both the EWN project and the plWordNet project had also goals other than the construction of a wordnet (consider the alignment of wordnets in the former, and automatic methods in the latter). Still, it is more appropriate than a comparison to wordnets developed over a much longer period. The relation between plWordNet and the contemporary wordnets is briefly described at the end of this section.

	Czech WN		Estonian WN		French WN		German WN	
	Synsets	LUs	Synsets	LUs	Synsets	LUs	Synsets	LUs
Nouns	9727	13829	5028	8226	17826	24499	9951	13656
Verbs	3097	6120	2650	5613	4919	8310	5166	6778
All	12824	19949	7678	13839	22745	32809	15132	20453

Table 5.2: Selected counts for Czech, Estonian and German wordnets built in the second phase of the EuroWordNet project (Vossen et al., 1999, p. 7)

Facts about the Czech, Estonian, French and German wordnets, built in the second phase of the EWN project, appear in Table 5.2. The nominal part of plWordNet is larger than in the Estonian wordnet, similar in size to the Czech and German wordnets, and significantly smaller only than the French wordnet, whose construction was based on an extensive translation of PWN. The verbal part of plWordNet is smaller than in any of the EWN wordnets. The reason is that the semi-automatic expansion was performed mainly on the nominal part. In 1.51 person-months, we added 60% lemmas and 40% LUs; every element was added and verified by two linguists. The EWN wordnets did not, in practice, include LUs other than nominal and verbal units, while plWordNet contains 3881 adjectival LUs. It should be emphasized, however, that all EWN wordnets are aligned with PWN 1.5 (via a subset of synsets used as a form of inter-lingua). Alignment with PWN was not a research goal in the plWordNet project.

	Including Monosemous Lemmas		Excluding Monosemous Lemmas	
	plWordNet	PWN 3.0	plWordNet	PWN 3.0
Nouns	1.317	1.24	2.361	2.79
Verbs	1.286	2.17	2.390	3.57
Adjectives	1.472	1.40	2.749	2.71

Table 5.3: Average polysemy in plWordNet and PWN 3.0

Polysemy rates for plWordNet in comparison to the rates of PWN 3.0 appear in Table 5.3. The adjectival part has not been semi-automatically expanded and represents the state from the core plWordNet. Both adjectival polysemy rates in plWordNet are similar to those of PWN, but it is hard to draw any general conclusions: plWordNet is so much smaller than PWN, and it only underwent a partial semi-automatic expansion.

	Average number of LUs per one synset				
	Czech WN	Estonian WN	French WN	German WN	plWordNet
Nouns	1.42	1.64	1.37	1.37	1.36
Verbs	1.98	2.12	1.69	1.31	2.42

Table 5.4: The average number of LUs per synset in plWordNet and four EWN wordnets (second phase) (Vossen et al., 1999, p. 7).

The definition of a synset in plWordNet, based on linguistic criteria (Section 2.1), may – on the face of it – lead to very small synsets, most of them with just one LU. Encouragingly, then, the average number of LUs per synset in the nominal part of plWordNet is 1.36. This number is very close to those obtained for the three largest EWN wordnets – see Table 5.4. The ratio for verbal synsets in plWordNet is rather high, but that part has been expanded in a small degree. We can expect a decrease in the ratio of the verbal LUs per synset during later stages of semi-automatic expansion. For the nominal part of plWordNet, the ratio decreased from 3.13 in version 12.2006 and 1.36 in the version 1.0.

	Percentage of synsets including the n LUs [%]									
	1	2	3	4	5	6	7	8	9	≥ 10
Nouns	79.50	12.67	4.22	1.84	0.91	0.34	0.20	0.12	0.07	0.13
Verbs	27.31	36.34	19.89	8.39	4.35	1.67	0.97	0.43	0.27	0.38
Adjectives	56.90	23.33	11.76	3.84	2.31	0.65	0.60	0.23	0.14	0.24

Table 5.5: Synset sizes.

In Table 5.5 we take a closer look at the distribution of the synset sizes over the three parts of plWordNet. Each number in the table shows what percentage of synsets belonging to the given part – nominal, verbal, adjectival – is such that they include the particular number of LUs. For the nominal part, the majority of synsets are singletons, but there is a significant percentage of 2- and 3-element synsets. The numbers gradually decrease with the increasing n. As a results of plWordNet expansion, many new non-singleton synsets were created or complemented with new LUs.

The largest existing nominal synset includes emotionally marked nominal LUs:

{*głupiec* 'fool', *głąb* 'noodle', *baran* 'blockhead', *osioł* 'donkey', *gamoń* 'nincompoop', *matoł* 'nitwit', *ćwierćinteligent* ‘pseudo-intellectual’, *bęcwał* ‘dolt’, *głupol* ‘dunderhead’, *głupi* ‘stupid’, *głupek* ‘booby’, *idiota* ‘idiot’, *przymuł* ‘(no equivalent)’, *muł* ‘mule’ ,*przygłup* ‘half-wit’, *cymbał* ‘chuckle-head’, *ciemna masa* ‘rabble’, *trąba* ‘dummy’, *imbecyl* ‘imbecile’, *tuman* ‘twit’, *bałwan* ‘dimwit’, *półgłówek* ‘blockhead’, *błazen* ‘fool’, *debil* ‘retard’, *dureń* ‘bonehead’, *klown* ‘clown’, *kretyn* ‘cretin’, *pajac* ‘buffoon’, *analfabeta* ‘illiterate’, *cep* ‘lunkhead’}[5]

In the case of the verbal part, the largest created synset contains a lot metaphorical descriptions:

{ *umrzeć* ‘die’, *skonać* ‘perish’, *dokonać żywota* ‘end one’s days’, *dokonać życia* ‘end one’s life’, *wyzionąć ducha* ‘give up the ghost’, *rozstać się z życiem* ‘part with one’s life’, *pożegnać się z życiem* ‘bid farewell to one’s life’, *pożegnać się ze światem* ‘bid farewell to the world’, *zakończyć życie* ‘end one’s life’, *przenieść się na łono Abrahama* ‘move to the bosom of Abraham’, *zemrzeć* ‘die’, *przenieść się na tamten świat* ‘move to the other world’, *przenieść się do wieczności* ‘move to eternity’, *przenieść się do lepszego świata* ‘move to the better world’, *zasnąć na wieki* ‘go to sleep forever’, *zasnąć snem wiecznym* ‘sleep eternal sleep’}

Finally, the largest adjectival synset contains highly emotionally marked adjectives:

{*fantastyczny* ‘fantastic’, *niepospolity* ‘outstanding’, *rewelacyjny* ‘sensational’, *zdumiewający* ‘amazing’, *wyjątkowy* ‘exceptional’, *niesamowity* ‘incredible’, *świetny* ‘splendid’, *niezwykły* ‘unusual’, *duży* ‘great’, *znakomity* ‘superb’, *kapitalny* ‘brilliant’, *wspaniały* ‘magnificent’, *wyśmienity* ‘≈excellent’}

All three largest synsets represent specific types of the language usage. The nominal and adjectival include LUs of very imprecise meaning, conveying more emotional meaning then descriptive. The verbal one groups LUs of quite precise meaning, but refers the topic that people avoid naming directly.

However, when we take a look at several smaller, but still large, nominal synsets, we can notice that their construction is not based on as simple rule as the largest synsets, for example:

[5] This group can be translated into English in hundreds of ways. What you see is an educated guess.

{*bok* 'side', *krawędź* 'edge', *skraj* 'brink', *kraj* 'brink', *kant* 'edge', *brzeg* 'margin', *obrzeże* 'margin'}

{*dobra strona* 'good side', *plus* 'plus', *cnota* 'virtue', *walor* 'value', *pozytyw* 'positive', *przymiot* 'attribute', *wartość* 'value', *zaleta* 'advantage'}

{*finanse* 'finances', *fundusz* 'fund', *kapitał* 'capital', *budżet* 'budget', *środki finansowe* 'financial means', *fundusze* 'funds'}

{*grób* 'grave', *mogiła* 'grave', *grobowiec* 'thomb', *nagrobek* 'gravestone', *miejsce pochówku* 'place of burial'}

{*istota* 'essence', *sens* 'sens', *sedno* 'core', *główne zagadnienie* 'main issue', *meritum* 'crux', *kwintensencja* 'quintessence', *jądro* 'gist'}

{*nierozdzielność* 'inseparability', *nierozerwalność* 'indissolubility', *jednolitość* 'uniformity', *spoistość* 'cohesiveness', *nierozłączność* 'inseparability', *jedność* 'unity', *spójność* 'cohession'}

The verbal and adjectival synsets are more diverse in size (Table 5.5), but it should be emphasised that the verbal part has been only expanded a little, and the adjectival part is the same as in the core plWordNet. The numbers of synsets and LUs are much smaller than for the nominal part, so a lot of the more specific LUs have not been added yet.

	Percentage of lemmas belonging to the n synsets [%]									
	1	2	3	4	5	6	7	8	9	≥ 10
Nouns	76.70	17.68	3.88	1.08	0.38	0.18	0.07	0.03	0.00	0.00
Verbs	79.41	14.87	4.03	1.23	0.29	0.17	0.00	0.00	0.00	0.00
Adjectives	72.99	15.90	6.56	2.66	1.10	0.08	0.23	0.15	0.15	0.18

Table 5.6: The number of synsets to which a lemma belongs

Table 5.6 presents a more detailed picture of the lemma polysemy. The numbers of monosemous lemmas appear in Table 5.3. Here they are expressed as the percentages in the first column. It is worth noticing that the percentage of monosemous nominal lemmas is lower than for the other two categories, in contrast with the much higher percentage of singleton synsets. It means that a large number of nominal lemmas are described by several singleton synsets. The columns from 2nd to the 10th describe polysemous lemmas. Observe a slight tendency of verbal and adjectival lemmas to have more meanings than do the nominal ones.

Finally, Table 5.7 presents the number of instances of each lexico-semantic relation. Some relations are defined only for particular parts of speech. The table cells for the undefined combinations are filled with ‘—’. As expected, the derivational relations dominate in the verbal and adjectival parts of plWordNet. For verbs, the set of defined relations is clearly not rich enough. We plan to expand it in the future.

Relation	No. instances			
	Nouns	Verbs	Adjectives	*All*
Hypernymy	12150	687	155	12992
Holonymy	1454	0	0	1454
Meronymy	1563	0	0	1563
Troponymy	0	37	0	37
Antonymy	1212	173	1618	3003
Conversion	35	66	0	101
Relatedness	981	2618	1226	4825
Pertainymy	1469	191	295	1955
Fuzzynimy	640	44	423	1107

Table 5.7: The number of lexico-semantic relations in plWordNet

Hypernymy among nouns seems to tend toward depth (perhaps due to the definition of the synset, assumed in plWordNet), but the limited size of the plWordNet 1.0 does not allow fully justified conclusions yet. The longest hyponymy path has 11 links:

{*istota* ‘being’}

→{*człowiek* ‘human’, *istota ludzka* ‘human being’, *homo sapiens* ‘Homo sapiens’, *człowiek rozumny* ‘sapient (human)’}

→{*człowiek ze względu na swoje zajęcie* ‘human by occupation’}

→{*pracownik* ‘employee’}

→{*pracownik ze względu na rodzaj pracy* ‘employee by job’}

→{*pracownik instytucji publicznych* ‘public servant’}

→{*funkcjonariusz* ‘officer’}

→{*żołnierz* ‘soldier’}

→{*komandos* ‘commando, ranger’}

→{*spadochroniarz* ‘paratrooper’, *skoczek spadochronowy* ‘parachutists’}

The semi-automatic expansion of plWordNet focused mainly on the nominal part. 650 nominal synsets – a relatively high number – have been linked to more than one

hypernym. The maximum of 4 hypernyms has been reached in the synset
{*mróz* 'frost', *zimnica* '≈cold', *ziąb* 'cold'}
which is a hyponym of
{*zjawisko atmosferyczne* 'atmospheric phenomenon'},
{*zimno* '≈cold'},
{*brak ciepła* 'lack of warm'} – an artificial LU,
{*niepogoda* 'bad weather'}.

The same phenomenon can also be observed among verbs and adjectives. 25 verbal synsets and 32 adjectival synsets have more than one hypernym.

5.3 Lessons Learned

When the plWordNet project began, no ready-to-use software to support wordnet construction was available. The first order of business, therefore, was to design and implement such a tool. In the meantime, manual work on the core plWordNet had to be initiated. In effect, we tackled three problems at once:

- the construction of a wordnet from scratch (initially without systematic software support),
- the design of support tools (initially without much insight and feedback of the eventual users, the linguists),
- and the theoretical and practical issues of semi-automated acquisition of material for wordnet development.

The first step in the manual construction of plWordNet was the acquisition of a list of lemmas to make up the core set of LUs. This step might be unproblematic given a large, sufficiently balanced corpus. We worked with the largest corpus of Polish available at that time, the IPI PAN Corpus, version 1.0 [IPIC] (Przepiórkowski, 2004). While no corpus can be ideally balanced, IPIC 1.0 covered a good variety of genres (Przepiórkowski, 2006) and was large enough – about 254 million tokens – so the most frequent lemmas in it should have been quite reasonable. As discussed in Section 2.4, however, there are difficulties worth paying attention to: tagging and preprocessing inaccuracies affect the lemma list, and so does the absence of many general LUs to "feed" the top part of the hypernymy hierarchy. The latter speaks strongly against sticking to a frequency list. It is better to compile manually or otherwise acquire a list of lemmas from a small unilingual or bilingual dictionary. Good lexicographic practice ensures that the most important lemmas are present, while the small size of a dictionary naturally limits the vocabulary considered.

In the end, we took the dictionary of Piotrowski and Saloni (1999) to complete the set of lemmas for the core plWordNet. This was necessary during the semi-automatic expansion of the nominal part of plWordNet (Section 4.5.4). We also briefly experimented with translating into Polish LUs from the upper levels of the nominal hypernymy structure in PWN. This resulted in many artificial LUs – unlexicalised meaning descriptions. We were more successful with a simple kind of "machine translation" applied to the verbal and adjectival parts of PWN. For each lemma occurring in the preselected part of PWN, we added to the list all its translations found in the electronic version of the dictionary of Piotrowski and Saloni (1999). We did not try to disambiguate translations, because we were interested in completing the list of lemmas, which was further processed manually. Moreover, the dictionary is a small pocket dictionary, so it should contain only the most general lemmas, and thus it acts as a kind of filter which eliminates translations for all infrequent LUs.

The weakness of the initial lemma list was problematic when expanding the core plWordNet via the WordNet Weaver (Sections 4.5.3 and 4.5.4): missing hypernyms – LUs from the upper parts of the hypernym structure – affected the accuracy of the algorithm.

The assumption that the LU is the centrepiece of plWordNet (Section 2.1) was not uncontroversial: synsets are key elements in most applications of a wordnet. The resulting structure, however, is not without advantages. The assumption is well motivated by the linguistic tradition and the lexicographic practice. The rules for adding new LUs, instances of semantic relations and, especially, LUs to synsets were systematically defined following the established linguistic tradition, and implemented in the plWordNetApp application (Section 2.4) as automatically filled substitution tests.

Synonymy is an elusive relation, not easily defined, yet it underlies the central notion of a synset. The construction of synsets in many wordnets is thus based on imprecise rules and on references to the extralinguistic properties of LUs. In plWordNet, a synset is defined through the lexico-semantic relations among its members; more precisely, it is the other way round – the similarity of several LUs due to the shared set of lexico-semantic relation targets[6] makes them candidates for the same synset. In the identification of synsets, hypernymy and meronymy have been distinguished as defining the structure (Section 2.1). This means that plWordNet (and generally any wordnet designed according to our method of defining synsets) is a network of LUs connected by lexico-semantic relations. A synset is in this case just a "shortcut" for the fact that two or more LUs share the same relations. The structure of plWordNet is based on the lexico-semantic relations among LUs, which are well established in linguistics and for which substitution tests are well known, so the linguists are likely to make highly consistent decisions.

[6] For a LU x, all LUs in some lexico-semantic relation with x are such targets.

The assumption that LUs in the same synset are semantically very close and can be mutually exchanged *in context* is important in applications of a wordnet. In plWordNet, with its much smaller synsets, such exchange works very well, because synsets are defined according to strict rules. There are fewer opportunities of exchange, because synsets are on average smaller (Section 5.2), but one can explore the hypernymy structure and "fold" direct hyponyms and hypernyms into groups of highly semantically similar LUs. The upside of this situation must be investigated experimentally.

Absolute synonymy is extremely rare in natural language. The ratio of non-singleton to singleton synsets in the nominal part is smaller than in other wordnets (Table 5.4), but the number of non-singleton synsets continually grows during semi-automatic expansion of plWordNet. New LUs are typically added to existing synsets based only on the hypernymy structure. Non-singleton synsets in plWordNet represent a rather strict variant of synonymy. Here are several examples:

{*dekoncentracja* 'deconcentration', *nieuwaga* 'inattention', *brak koncentracji* 'lack of concentration', *rozproszenie* '≈being distracted'}
{*nakrętka* 'screw-top', *zakrętka* 'cap'}
{*sworzeń* 'pivot', *bolec* 'bolt', *trzpień* 'mandrel'}
{*pieczątka* 'stamp', *pieczęć* '(large) stamp', *stempel* 'stamp'}
{*afisz* 'placard', *plakat* 'poster'}
{*pokwitowanie* 'receipt', *kwit* 'receipt'}
{*rubryka* 'column', *kolumna* 'column', *łam* 'column'}
{*grafika* 'graphics', *rysunek* 'drawing', *rycina* 'drawing'}

We assumed at first that it would be possible to create a core plWordNet which consists only of LUs fully described by all appropriate instances of lexico-semantic relations. It has turned out that it was not quite feasible to follow this assumption. LUs are defined by sets of lexico-semantic relation instances. A truly complete description of several thousand core LUs would require another several thousand LUs to define the relation links in full. Those LUs, in turn, would require more units for comprehensive description, and so forth. In addition, limiting the lexical material to "general language" introduced undesirable vagueness, and in any event it seems inevitable that omissions will creep in. A more practical goal is now to try to include in the core plWordNet all general LUs located high in the hypernymy structure. A core plWordNet that respects this assumption would be a good starting point for semi-automatic expansion next, but it is hard to formulate a prescription that would lead to such a result – other than starting from the entry list of a small dictionary.

We have tested many methods of extracting lexico-semantic relations (Chapters 3 and 4). None of them ensures quality comparable to manual work. While the accuracy was often good, the problem was to find the limits. For example, Measures of Semantic Relatedness [MSR] produce continuous results. Defined for any pair of lemmas, the extraction based on lexico-syntactic patterns produces LU pairs that represents different shades of some semantic relation. It is very hard to construct a general automatic mechanism that defines the border between those potential relation instances which are correlated with the linguist's judgment, and those which are not. It becomes easier when we consider automatic expansion of an *existing* wordnet. The linguistic knowledge already represented by the wordnet structures helps increase the trustworthiness of the automated additions. Consider the promising results of the WordNet Weaver application (Section 4.5.4).

A core plWordNet should contain the upper levels of the hypernymy hierarchy, but it is very hard to construct it top-down without compromising the linguistic nature of the lexical network: one can unwittingly "slip" into an abstract ontology (taxonomy). More general LUs have few true hypernyms, and it is difficult to distinguish between their direct and indirect hyponyms. Bottom-up work might be safer, but it too has a drawback: the proper selection[7] of the more specific LUs in order to "activate" a wide range of more general LUs at the end of the process of core plWordNet construction. The problem is to make the selection in such a way that we can get an exhaustive set of the most general LUs if we just keep describing hypernyms of the specific LUs. Our experience suggests strong preference for the bottom-up approach. It worked especially well during the semi-automatic expansion phase.

Our experiments with the WordNet Weaver [WNW], a tool for semi-automatic wordnet expansion (Sections 4.5.3-4.5.4), were generally encouraging. 8361 new lemmas, 10537 new LUs, 8729 synsets and 11063 instances of lexico-semantic relations have been added to the core plWordNet at the cost of 3.4 person-months. Every decision assisted by WNW was verified by a coordinator, and many improvements were made to the initial plWordNet structure.

It is hard to separate the time spent on correcting the core plWordNet from the time spent on expanding it. WNW allowed us to discover many errors in the core plWordNet structure: trying to attach new lemmas to the existing structure often brought out the drawbacks of that structure. WNW's suggestions are less helpful for gerunds (it is an open problem how to reconcile gerund description by both MSRs and by pattern). The percentage of usefully suggested attachment varies across domains.

We initially assumed a model of first generating part of the plWordNet structure and then correcting it semi-automatically. That did not work well at all. A linguist could be lost if required to make hundreds of corrections in a continuously evolving wordnet structure. It seems better to correct the proposed attachments one by one.

[7] Such selection is necessarily constrained, not the least by financial considerations.

We clustered automatically new lemmas processed by WNW. That appeared to help: clusters were small, so as not to overwhelm the linguist, but at the same time each cluster was organised around one to three domains. Often better suggestions arose from processing some lemmas from a cluster and running the algorithm of Activation-area Attachment [AAA] (Section 4.5.3) on the rest.

In spite of investing a lot of effort in different approaches to clustering documents and lemmas for extracting synsets and discovering LUs, automatic extraction of complete synsets and structures built of synsets is an unsolved problem if it is to come close to human performance. According to our definition of a synset, however, it should be possible after more research, because the basis is the identification of the shared targets of semantic relations.

AAA combines the results generated by several extraction methods. Each method alone gives subpar results, too noisy to be useful for the linguists. The combination, however, gave a positive outcome. Information on lexico-semantic relations in corpora is ambiguous and partial, so it seems natural to try to combines sources.

While semi-automated approach is feasible, it requires a very large corpus in order to get good results on general vocabulary. For the final experiments, we applied a joint corpus of about 581 million tokens (Section 3.4.5). The experiments performed with MSR extraction and pattern-based extraction on different combinations of parts of the joint corpus (Section 4.3) showed that accuracy grows continuously with the increasing size of the subcorpus. It is hard to identify any point at which the process stabilises. The upside is that we successfully made do without advanced language tools. We worked only with a morphosyntactic tagger and morphosyntactic constraints based on the tagger engine.

As expected, WNW does not work well for new lemmas which have a general meaning and should be located high in the hypernymy structure. The suggested attachments are usually limited to more or less direct hyponyms. AAA cannot combine these substructures since the connection point is just lacking. Moreover, the evidence is not evenly distributed over the different LUs in plWordNet, so the proposed attachment can sometimes appear accidental.

The work on the semi-automatic expansion of plWordNet focussed on nominal hypernymy. There are two reasons: most techniques proposed in the literature have been tried on nouns, and pattern-based methods apply almost exclusively to nominal LUs. We did perform the first experiments with the application of AAA to verbal LUs using only an MSR for verbs as the knowledge source. The results are promising, so it appears possible to go beyond nouns. It is worth noting that now, after having constructed the expanded version of plWordNet, we have the knowledge sources comparable to those which made it possible for Girju et al. (2006) to achieve good results for meronymy. There is no reason to expect that we will not succeed in extracting meronymy, at least if we follow their approach.

The derivational relations are an important part of the whole network. Instances of derivational relations are often the only way of describing the meaning of adjectival LUs. For future applications of plWordNet it is necessary to develop a more fine-grained semantic description of derivationally motivated links.

Much effort went into research on the various methods of automatic extraction of lexico-semantic relations and on the construction of the WNW system. A wordnet of a similar size, perhaps larger, might be constructed at the same cost without the support of automatic tools. The technology which we created, however, is ready to use; it opens interesting possibilities for further improvements and extensions. The latest version of WNW (Section 4.5.4) was used consistently for several months. That effort brought about a significant increase in the number of lemmas, LUs and instances of lexico-semantic relations. WNW also helps find inconsistencies and missing elements in the plWordNet structure.

A final point: much of the previous work on wordnet development has focussed on English, a language which typologically differs significantly from Polish. No research similar to our appears to have been published for other Slavic language. The lack of applicable comparable experience added to the workload.

5.4 What Next?

It has been not so much the project deadline as the depletion of funds that has prompted us to announce the completion of plWordNet, version 1.0. Our plans go much further, naturally. We perceive the construction of plWordNet as a continuous process. Its end cannot be easily predicted, given the amount of work – not to mention the fact that a *complete* wordnet is a moving target.

We will gradually and systematically add new LUs to plWordNet using the WordNet Weaver system. The system of relations among verbs will be further developed. It is an open question whether to expand it inside plWordNet, for example by describing subcategorisation frames and semantic selectional preferences in association with nouns in plWordNet, or to combine it with some other resource oriented toward the description of verbs.

We must begin to align plWordNet with other wordnets – PWN, GermaNet or BalkaNet are likely counterparts. We will also work toward making plWordNet a part of the Global WordNet Grid (Fellbaum and Vossen, 2007). Sometimes flying, sometimes walking, but always steadily moving towards a large wordnet for Polish: this is our plan. The pace, as expected, depends on funding.

Other than the obvious goal of "growing" plWordNet, we see two directions of future work:

- evaluation of the concept of linguistically motivated wordnet structure,
- further development of the algorithms of automatic extraction of semantic relations and the methods of semi-automated wordnet construction.

It may not be enough to justify the assumed wordnet model by analytical consideration, for example in comparison to the psychologically oriented concept of the Princeton WordNet. Sections 2 and 5.3 offer some discursive support. What is needed, clearly, is practice. Several Polish universities have been granted free research licences. The plWordNet web pages (`plwordnet.pwr.wroc.pl`) have had about 12000 visitors (based on unique IP addresses, more than 180000 visits). The real test is yet to come: a range of experiments in various applications of plWordNet.

We made a first step ourselves: we ran several Word Sense Disambiguation algorithms on Polish using plWordNet (Baś, 2008, Baś et al., 2008). 13 lemmas representing 54 LUs altogether were selected in such a way that the subsequent lemmas pose different problems with respect to hyponymy and polysemy. A small training/test subcorpus was collected, including sentences which represent different senses of the lemmas. The results are very promising in spite of the fine-grained sense distinction observed for several lemmas. Much more is needed. We plan to work on plWordNet, and we will actively publicise the system. We offer free research licences to anyone who has a research plan that includes plWordNet.

Our main wordnet development tool, the WordNet Weaver [WNW], works only with the hypernymy structure. It allows for editing synsets and hypernymic links while adding new lemmas to plWordNet. The hypernymy structure is necessarily shallower for adjectival and verbal LUs, so one should leverage all types of links between synsets and LUs in order to collect evidence for the most appropriate attachment point. Lexico-semantic relations other than hypernymy can also be beneficial for expanding the nominal part of plWordNet.

In WNW, any change in the lexico-semantic relations other than hypernymy is possible but from the main plWordNetApp, not from the WNW graphical browser. The algorithm of Activation-area Attachment [AAA] very often selects holonyms as possible attachment points. All this is especially limiting for verbs, and makes adding adjectival LUs almost impossible. We plan to enable editing of all types of lexico-semantic relation via WNW graphs.

The present model behind the AAA is heuristic. We plan to investigate its possible generalisations on the basis of the statistical properties of the different evidence and relation graph properties.

Besides WNW, there are also open research questions concerning the work of its components. There is no visible threshold in the values produced by the proposed Measures of Semantic Relatedness [MSR] based on the Rank Weight Function which distinguishes closely related lemmas from other lemmas. We plan to explore the statistical properties of the features in order to construct a kind of confidence measure which describes the evidence for the calculations for a particular lemma pair.

The accuracy of MSR increases with the increasing corpus size. We will keep collecting large corpora. The description based on the occurrences of the lexico-morphosyntactic constraints, however, is collected only from a relatively small percentage of occurrences of the lemmas in focus. Types of description other than the constraints can increase the utilisation ratio of the corpus. In the case of the Estratto algorithm (a pattern-based approach), the usage of more complex pattern scheme significantly improved the accuracy, so we want continue this line of development. An obvious extension for the approach based on classifiers presented in Section 4.5.1 is its application to other types of lexico-semantic relation such as holonymy/meronymy.

The most attractive aspect of the research is that opening one door reveals many other doors which are still closed.

The end? No! To be continued...

Appendix A

Tests for Lexico-semantic Relations

Synonymy

Test for nouns (T. I)

1.	If it is X, it is also Y	If it is a drink, it is also a beverage
2.	If it is Y, it is also X	If it is a beverage, it is also a drink

Test for verbs (T. II)

1.	If someone is doing X, they are also doing Y	If someone is swearing, they are is also cursing
2.	If someone is doing Y, they are also doing X	If someone is cursing, they are also swearing

Test for adjectives (T. III)

1.	Someone/something X is also Y	Someone/something stupid is also foolish
2.	Someone/something Y is also X	Someone/something foolish is also stupid

Antonymy

Test for nouns (T. IV)

1.	X and Y are kinds of Z	Man and woman are kinds of human being
2.	It is untrue that someone/something is not X and not Y at the same time	It is untrue that someone is not a man and not a woman at the same time

Test for verbs (T. V)

1.	X and Y are kinds of actions	
2.	If someone is doing X, they are not doing Y at the same time	If someone is pouring something in, they are not pouring it out
3.	If someone is doing Y, they are not doing X at the same time	If someone is pouring something out, they are not pouring it in

Test for adjectives (T. VI)

1.	Contradictory and complementary LUs	
a.	It is untrue that someone/something is X and Y at the same time	It is untrue that someone/something is alive and dead at the same time
b.	It is untrue that someone/something is not X and not Y at the same time	It is untrue that someone/something is not alive and not dead at the same time
2.	Opposite LUs	
a.	It is untrue that someone/something is X and Y at the same time	It is untrue that someone/something is long and short at the same time
b.	It is true that someone/something is not X and not Y at the same time	It is true that someone/something is not long and not short at the same time

Conversion

Test for nouns (T. VII)

1.	p1 and p2 are nouns	*Husband* and *wife* are nouns
2.	If X is Y's p1	If X is Y's husband
3.	then Y is X's p2	then Y is X's wife

Test for verbs (T. VIII)

1.	p1 and p2 are predicates with at least two arguments	
2.	If X p1 Y	If X is buying something from Y
3.	then Y p2 X	then Y is selling something to X

Test for adjectives (T. IX)

1.	p1 and p2 are adjectives in the comparative degree	
2.	If X is p1 than Y	If X is shorter than Y
3.	then Y is p2 than X	then Y is taller than X

Hypernymy/hyponymy

Test for nouns (T. X)

1.	X is a kind of Y (with certain features)	A car is a kind of a vehicle
2.	Y is not a kind of X	A vehicle is not a kind of a car
Z ($\neq$ X) is also a kind of Y	A bicycle is also a kind of a vehicle	

Test for verbs (T. XI)

1.	If someone is doing X, they are also doing Y	If someone runs, they also move
2.	If someone is doing Y, they are not necessary doing X	If someone moves, they do not necessary run
3.	There are other such activities	If someone walks, they also move

Test for adjectives (T. XII)

1.	Someone/something X is also Y	Someone/something scarlet is also red
2.	Someone/something Y is not necessary X	Someone/something red is not necessary scarlet
3.	Someone/something Z ($\neq$ X) is also Y	Someone/something carmine is also red

Meronymy/holonymy

Test for nouns (T. XIII)

1.	X is	
a.	a part of Y	*A wheel* is a part of a *bicycle*
b.	a portion Y	*A mouthful* is a portion of *juice*
c.	a place in (on) Y	*A backwoods* is a place in a *forest*
d.	an element of Y	*A book* is an element of a *library*
e.	stuff of which something Y is made	*Metal* is stuff of which a *blade* is made

2.	Y is a whole	
a.	whose parts are X . . .	*A bicycle* is a whole whose parts are *wheels*
b.	whose portion is X	*Juice* is a whole whose portion is a *mouthful*
c.	which is a place where X is	A *forest* is a place where *backwoods* is
d.	whose element is X	*A library* is a whole whose element is a *book*
e.	made (entirely or partially) of X	*A* blade is made of *metal*

Troponymy

Test for verbs (T. XIV)

1.	To X is to Y in a certain way	To *limp* is to *walk* in a certain way To *read till the end* is to *read* in a certain way
2.	To Y is not necessarily to X	To *walk* is not necessarily to *limp* To *read* is not necessarily to *read till the end*

Relatedness

Test for nouns (T. XV)

1.	X_{gerund} is derived from Y_{verb}	*reading* ← *read*
2.	X_{noun} is derived from $Y_{adjective}$	 *height* ← *high* *blackness* ← *black*

Test for verbs (T. XVI)

1.	Y is a perfective form of X	*have painted* ‘pomalować’ ← *paint (a wall)* ‘malować’ *have painted* ‘namalować’ ← *paint (a picture)* ‘malować’
2.	X_{verb} is to cause to be $Y_{adjective}$	*sadden* is to cause to be *sad*

Test for adjectives (T. XVII)

1.	$X_{adjective}$ is derived from Y_{noun}	*school*$_{attributive}$ ‘szkolny’ ← *school* ‘szkoła’
2.	$X_{participle}$ is derived from Y_{verb}	
a.	X refers to an ongoing activity	*reading*$_{participle}$ ‘czytający’ ← *read* ‘czytać’
b.	X refers to a state or a completed activity	*read*$_{past\ participle}$ ‘czytany’ ← *read* ‘czytać’

Pertainymy

Test for nouns (T. XVIII)

1.	Y_{noun} is such that	
a.	$X_{verb,\ adjective\ or\ noun}$ is an attribute of Y	*beard* ‘broda’ ← *bearded (man)* ‘brodacz’, *strange* ‘dziwny’ ← *someone strange (eccentric)* ‘dziwak’
b.	Y is a place for doing X_{verb}	*assembly* ‘montować’ ← *assembly room* ‘montownia’
2.	Y is what X does	*a cleaning person* ‘czyściciel’ ← *clean* ‘czyścić’
3.	X_{noun} can be paraphrased as female Y_{noun}	*painter (feminine)* ‘malarka’ ← *painter (masculine)* ‘malarz’, *mare* ← *stallion*
4.	X_{noun} can be paraphrased as young, juvenile Y_{noun}	*kitten* ‘←’ *cat* ‘,’ *foal* ‘←’ *horse* ‘
’		
5.	X_{noun} can be paraphrased as	
a.	small/large Y_{noun}	*large dog* ‘psisko’ ← *dog* ‘pies’, *doggy* ‘piesek’ ← *dog* ‘pies’
b.	expressing positive/negative feelings about Y_{noun}	*small child* ‘dziecinka’ ← *child* ‘dziecko’, *little (poor) woman* ‘kobiecina’ ← *woman* ‘kobieta’
6.	X_{noun} means a citizen of Y_{noun} or someone living in Y_{noun}	*Pole* ← *Poland*, *Texan* ← *Texas*

Test for adjectives (T. XIX)

1.	X is Y with a positive emotional attitude	*tiny* ‘maleńki, malutki’ ← *small* ‘mały’

Bibliography

Agarwal, Abhaya and Alon Lavie. (2008) "Meteor, m-bleu and m-ter: Evaluation metrics for high-correlation with human rankings of machine translation output". *Proceedings of the Workshop on Statistical Machine Translation at the 46th Annual Meeting of the Association of Computational Linguistics*. Columbus, OH, 115–118.

Agichtein, Eugène and Luis Gravano. (2000) "Snowball: extracting relations from large plain-text collections". *Proceedings of the Fifth ACM Conference on Digital Libraries*. New York, NY, USA: ACM, 85–94.

Agichtein, Eugene, Luis Gravano, Jeff Pavel, Viktoriya Sokolova and Aleksandr Voskoboynik. (2001) "Snowball: a prototype system for extracting relations from large text collections". *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*. New York, NY, USA: ACM, 612.

Agirre, Eneko and Philip Edmonds, ed. (2006) *Word sense disambiguation: Algorithms and applications*. Springer.

Alfonseca, Enrique and Suresh Manandhar. (2002) "Extending a lexical ontology by a combination of distributional semantics signatures". *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web – EKAW 2002*, 1–7.

Amsler, Robert A. (1981) "A taxonomy for English nouns and verbs". *Proceedings of the 19th Annual Meeting of the Association for Computational Linguistics*. Stanford, California, USA: Association for Computational Linguistics, 133–138.

Apresjan, Jurij D. (2000) *Semantyka leksykalna. Synonimiczne środki języka [Lexical semantics. Synonymy devices in language]*. Warszawa. Translated by Z. Kozłowska and A. Markowski.

Azarowa, Irina V. (2008) "RussNet as a computer lexicon for Russian". Kłopotek et al. (2008), 447–456.

Baś, Dominik. (2008) "Ujednoznacznianie sensu słów w polskim tekście [Word sense disambiguation in Polish text]". Master's thesis, Faculty of Computer Science and Management, Wroclaw University of Technology.

Baś, Dominik, Bartosz Broda and Maciej Piasecki. (2008) "Towards word sense disambiguation of Polish". *Proceedings of the International Multiconference on Computer Science and Information Technology – Third International Symposium Advances in Artificial Intelligence and Applications*, 73–78.

Bagga, Amit, Joyce Y. Chai and Alan W. Biermann. (1997) "The role of WordNet in the creation of a trainable message understanding system". *Proceedings of the 14th National Conference on Artificial Intelligence and the 9th Conference on the Innovative Applications of Artificial Intelligence*, 941–948. URL `http://www.cs.duke.edu/~amit/iaai97.ps.gz`

Banerjee, Satanjeev and Ted Pedersen. (2002) "An adapted Lesk Algorithm for word sense disambiguation using WordNet". *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics*, 136–145. URL `http://www.d.umn.edu/~tpederse/Pubs/cicling2002-b.pdf`

Basili, Roberto, Roberta Catizone, Lluís Padro, Maria Teresa Pazienza, German Rigau, Andrea Setzer, Nick Webb and Fabio Zanzotto. (2002) "Knowledge-based multilingual document analysis". *Proceedings of the COLING 2002 Workshop "SemaNet'02: Building and Using Semantic Networks"*. URL `http://www.cs.ust.hk/~hltc/semanet02/pdf/basili.pdf`

Bentivogli, Luisa, Pamela Forner, Bernardo Magnini and Emanuele Pianta. (2004) "Revising WordNet domains hierarchy: Semantics, coverage, and balancing". *Proceedings of the COLING 2004 Workshop on "Multilingual Linguistic Resources"*, 101–108.

Berland, Matthew and Eugene Charniak. (1999) "Finding parts in very large corpora". *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 57–64.

Berry, Michael W. (1992) "Large scale singular value computations". *International Journal of Supercomputer Applications* 6(1): 13–49.

Bloomfield, Leonard. (1933) *Language*. New York.

BNC. (2007) "The British National Corpus, version 3 (BNC XML Edition)". Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL `http://www.natcorp.ox.ac.uk/`

Boleda, Gemma, Toni Badia and Eloi Batlle. (2004) "Acquisition of semantic classes for adjectives from distributional evidence". *Proceedings of the 20th Conference of the International Committee on Computational Linguistics*. ACL, 1119–1125.

Boleda, Gemma, Toni Badia and Sabine Schulte im Walde. (2005) "Morphology vs. syntax in adjective class acquisition". *Proceedings of the ACL-SIGLEX Workshop* *on Deep Lexical Acquisition*. Ann Arbor, Michigan: Association for Computational Linguistics, 77–86.

Boyd-Graber, Jordan, Christiane Fellbaum, Daniel Osherson and Robert Schapire. (2006) “Adding dense, weighted, connections to WordNet”. Petr Sojka *et al.*, ed., *Proceedings of the Third International WordNet Conference*. Brno: Masaryk University, 29–36.

Brill, Eric. (1995) “Transformation-based error-driven learning and natural language processing: A case study of part-of-speech tagging”. *Computational Linguistics* 21(4): 543–566.

Brin, Sergey. (1999) “Extracting patterns and relations from the World Wide Web”. *WebDB ’98: Selected papers from the International Workshop on The World Wide Web and Databases*. London, UK: Springer-Verlag, 172–183.

Broda, Bartosz. (2007) “Mechanizmy grupowania dokumentów w automatycznej ekstrakcji sieci semantycznych dla języka polskiego [Mechanisms for grouping documents in automatic extraction of semantic networks for Polish]”. Master’s thesis, Wydział Informatyki i Zarządzania, Politechnika Wrocławska.

Broda, B., M. Derwojedowa and M. Piasecki. (2008) “Recognition of structured collocations in an inflective language”. *Systems Science* 34(4): 27–36 extended version of a paper in the Proceedings of AAIA’08, Wisła, Poland.

Broda, Bartosz, Magdalena Derwojedowa, Maciej Piasecki and Stanisław Szpakowicz. (2008) “Corpus-based semantic relatedness for the construction of Polish WordNet”. *Proceedings of the Sixth Conference on Language Resources and Evaluation*. URL `http://www.lrec-conf.org/proceedings/lrec2008/`

Broda, Bartosz and Maciej Piasecki. (2008a) “Experiments in documents clustering for the automatic acquisition of lexical semantic networks for Polish”. Kłopotek et al. (2008), 203–212.

———. (2008b) “SuperMatrix: a General Tool for Lexical Semantic Knowledge Acquisition”. Kłopotek et al. (2008), 345–352.

Broda, Bartosz, Maciej Piasecki and Stanisław Szpakowicz. (2008) “Sense-based clustering of polish nouns in the extraction of semantic relatedness”. *Proceedings of the International Multiconference on Computer Science and Information Technology – Second International Symposium* Advances in Artificial Intelligence and Applications, 83–89. URL `http://iis.ipipan.waw.pl/2008/proceedings.html`

———. (2009) “Rank-Based Transformation in measuring semantic relatedness”. *Canadian Conference on AI*, 187–190.

Budanitsky, Alexander and Graeme Hirst. (2006) “Evaluating WordNet-based measures of semantic distance”. *Computational Linguistics* 32(1): 13–47.

Bullon, Stephen, Chris Fox, Elizabeth Manning, Michael Murphy, Ruth Urbom and Karen Cleveland Marwick, ed. (2003) *Longman Dictionary of Contemporary English*. Pearson Education Limited. Fifth impression.

Calzolari, Nicoletta, Claire Cardie and Pierre Isabelle, ed. (2006) *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. The Association for Computer Linguistics.

Caraballo, Sharon A. (1999) "Automatic construction of a hypernym-labeled noun hierarchy from text". *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. Baltimore, MD, 120–126.

———. (2001) *Automatic construction of hypernym-labeled noun hierarchy from text*. Ph. D. dissertation, The Department of Computer Science, Brown University.

Ceglarek, Dariusz and Wojciech Rutkowski. (2006) "Automated acquisition of semantic relations for information retrieval systems". *Proceedings of the 10th International Conference on Business Information Systems*. Springer, 217–228.

Choi, Key-Sun and Hee-Sook Bae. (2004) "Procedures and problems in Korean-Chinese-Japanese WordNet with shared semantic hierarchy". Sojka et al. (2004), 91–96.

Clark, Peter, Christiane Fellbaum and Jerry Hobbs. (2008) "Using and extending WordNet to support question-answering". Tanács et al. (2008), 111–119.

Clough, Paul and Mark Stevenson. (2004) "Evaluating the contribution of EuroWordNet and word sense disambiguation to cross-language information retrieval". *Proceedings of the Second Global WordNet Conference*. Brno, Czech Republic, 97–105. URL `http://www.fi.muni.cz/gwc2004/proc/73.pdf`

Cohen, Jacob. (1960) "A coefficient of agreement for nominal scales". *Educational and Psychological Measurement* 20: 3–46.

Cramer, Irene and Marc Finthammer. (2008) "An evaluation procedure for WordNet based lexical chaining: Methods and issues". Tanács et al. (2008), 120–146.

Day, David, John Aberdeen, Lynette Hirschman, Robyn Kozierok, Patricia Robinson and Marc Vilain. (1997) "Mixed-initiative development of language processing systems". *Proceedings of the Fifth ACL Conference on Applied Natural Language Processing*, 348–355.

Debasri, Chakrabarti, Narayan Dipak Kumar, Pandey Prabhakar and Bhattacharyya Pushpak. (2002) "Experiences in building the IndoWordNet: a WordNet for Hindi". *Proceedings of the First Global WordNet Conference*. Mysore, India.

Dernowicz, Wiktor. (2007) "Automatic acquiring of semantic relations from text collections". Vetulani (2007), 310–314.

Derwojedowa, Magdalena, Maciej Piasecki, Stanisław Szpakowicz and Magdalena Zawisławska. (2008) "plWordNet – The Polish Wordnet". Online access to the database of plWordNet: `www.plwordnet.pwr.wroc.pl`.

Derwojedowa, Magdalena, Maciej Piasecki, Stanisław Szpakowicz, Magdalena Zawisławska and Bartosz Broda. (2008) "Words, concepts and relations in the construction of Polish WordNet". Tanács et al. (2008), 162–177.

Derwojedowa, Magdalena and Michał Rudolf. (2003) "Czy Burkina to dziewczyna i co o tym sądzą ich królewskie mości, czyli o jednostkach leksykalnych pewnego typu [... on certain type of lexical units]". *Poradnik Językowy* 5/2003: 37–49.

Derwojedowa, Magdalena, Stanisław Szpakowicz, Magdalena Zawisławska and Maciej Piasecki. (2008) "Lexical units as the centrepiece of a wordnet". Kłopotek et al. (2008), 351–357.

Derwojedowa, Magdalena and Magdalena Zawisławska. (2007a) "Relacje leksykalne w polskiej i czeskiej bazie wordnet [Lexical relations in the Polish and Czech wordnets]". *Z Polskich Studiów Slawistycznych* Seria XI, Językoznawstwo: 15–23.

———. (2007b) "Relacje leksykalne w polskiej i czeskiej bazie wordnet [Lexical relations in the Polish and Czech wordnets]". *Z Polskich Studiów Slawistycznych* Seria XI, Językoznawstwo: 15–23.

Derwojedowa, Magdalena, Magdalena Zawisławska, Maciej Piasecki and Stanisław Szpakowicz. (2007) "Relacje w polskim WordNecie [Relations in Polish WordNet]". Report, the PREPRINTY Series 1, Institute of Applied Informatics, Wrocław University of Technology. URL `http://plwordnet.pwr.wroc.pl/main/content/files/publications/relacje_v5rc02.pdf`

Dorr, Bonnie J. (1997) "Large-scale dictionary construction for foreign language tutoring and interlingual machine translation". *Machine Translation* 12(1): 1–55. URL `ftp://ftp.umiacs.umd.edu/pub/bonnie/icall-97.ps`

Dubisz, Stanisław, ed. (2004) *Uniwersalny słownik języka polskiego [Universal Dictionary of Polish Language], electronic version 0.1*. PWN.

Dunning, Ted. (1993) "Accurate methods for the statistics of surprise and coincidence". *Computational Linguistics* 191: 61–74.

Edmonds, Philip and Graeme Hirst. (2002) "Near-synonymy and lexical choice". *Computational Linguistics* 28(2): 105–144.

Fellbaum, Christiane. (1998a) "A semantic network of English: The mother of all WordNets". *Computers and the Humanities* 32: 209–220.

———. (1998b) *A semantic network of English verbs.*, Fellbaum (1998c)., chapter 3, 69–104.

———, ed. (1998c) *WordNet – an electronic lexical database*. The MIT Press.

Fellbaum, Christiane and Piek Vossen. (2007) “Connecting the universal to the specific: Towards the global grid”. Toru Ishida *et al.*, ed., *Intercultural Collaboration: First International Workshop*. New York: Springer, 1–16.

Firth, John Rupert. (1957) “A synopsis of linguistic theory 1930-55”. *Studies in Linguistic Analysis* (special volume of the Philological Society): 1–32 Oxford, The Philological Society.

Fleiss, Joseph L. (1971) “Measuring nominal scale agreement among many raters”. *Psychological Bulletin* 76(5): 378–382.

Forster, Richard. (2006) *Document clustering in large German corpora using natural language processing*. Ph. D. dissertation, University of Zurich.

Francis, Winthrop Nelson and Henry Kučera. (1982) *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin.

Freitag, Dayne, Matthias Blume, John Byrnes, Edmond Chow, Sadik Kapadia, Richard Rohwer and Zhiqiang Wang. (2005) “New experiments in distributional representations of synonymy”. *Proceedings of the Ninth Conference on Computational Natural Language Learning*. Ann Arbor, Michigan: Association for Computational Linguistics, 25–32.

Fromkin, Victoria, Bruce Hayes, Susan Curtiss, Anna Szabolcsi, Tim Stowell, Edward Stabler, Dominique Sportiche, Hilda Koopman, Patricia A. Keating, Pamela Munro, Nina Hyams and Donca Steriade. (2000) *Linguistics: An introduction to linguistic theory*. Blackwell Publishing.

Fukumoto, Fumiyo and Yoshimi Suzuki. (2001) “Learning lexical representation for text categorization”. *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, 156–161. URL `http://www.seas.smu.edu/~rada/mwnw/papers/WNW-NAACL-232.pdf.gz`

Geffet, Maayan and Ido Dagan. (2004) “Vector quality and distributional similarity”. *Proceedings of the 20th Conference of the International Committee on Computational Linguistics*, 247–254.

Girju, Roxana, Adriana Badulescu and Dan Moldovan. (2006) “Automatic discovery of part-whole relations”. *Computational Linguistics* 32(1): 83–135.

Gonzalo, Julio, Felisa Verdejo, Irina Chugur and Juan Cigarran. (1998) “Indexing with WordNet synsets can improve text retrieval”. *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, 38–44. URL `http://www.ai.sri.com/~harabagi/coling-acl98/acl_work/julio.ps.gz`

Graliński, Filip. (2005) “A simple CF formalism and free word order”. *Archives of Control Sciences* 15(LI)(3): 541–554.

Grefenstette, Gregory. (1993) “Evaluation techniques for automatic semantic extraction: Comparing syntactic and window based approaches”. *Proceedings of The Workshop on Acquisition of Lexical Knowledge from Text, SIGLEX’93*. ACL, 205–216.

Guha, Sudipto, Rajeev Rastogi and Kyuseok Shim. (2000) “ROCK: A Robust Clustering Algorithm for Categorical Attributes”. *Information Systems* 25(5): 345–366. URL `citeseer.ist.psu.edu/guha00rock.html`

GWA. (2008a) “The Global WordNet Association”. Homepage of the association. URL `http://www.globalwordnet.org/`

———. (2008b) “Wordnets in the world”. Web page maintained by the Global Wordnet Association. URL `http://www.globalwordnet.org/gwa/wordnet_table.htm`

Hamp, Birgit and Helmut Feldweg. (1997) “GermaNet – a Lexical-Semantic Net for German”. *Proceedings of the ACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid, 9–15.

Harris, Zellig Sabbetai. (1968) *Mathematical structures of language*. New York: Interscience Publishers.

Hatzivassiloglou, Vasileios and Kathleen R. McKeown. (1993) “Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning”. *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*. ACL, 172–182.

Hearst, Matti A. (1992) “Automatic acquisition of hyponyms from large text corpora”. *Proceedings of the Conference of the International Committee on Computational Linguistics*. Nantes, France: The Association for Computer Linguistics, 539–545.

Hearst, Marti A. (1998) *Automated discovery of WordNet relations*., Fellbaum (1998c)., chapter 5, 131–151.

Herrera, Jesús, Anselmo Peñas and Felisa Verdejo. (2006) “Textual entailment recognition based on dependency analysis and WordNet”. *First PASCAL Machine Learning Challenges Workshop*, 231–239.

Hindle, Donald. (1990) “Noun classification from predicate-argument structures”. *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*. Pittsburgh, PA: ACL, 268–275.

Hirst, Graeme and Alexander Budanitsky. (2005) “Correcting real-word spelling errors by restoring lexical cohesion”. *Natural Language Engineering* 11(01): 87–111.

Hockett, Charles F. (1964) *A course in modern linguistic*. Macmillan, New York.

Horák, Aleš, Karel Pala, Adam Rambousek and Martin Povolný. (2006) “DEBVisDic – first version of new client-server WordNet browsing and editing tool”. *Proceedings of the Third Global WordNet Conference*. Masaryk University, 325–328. URL http://nlp.fi.muni.cz/publications/gwc2006_hales_pala_etal/gwc2006_hales_pala_etal.pdf

Horák, Aleš and Pavel Smrž. (2004) "New features of WordNet editor VisDic". *Romanian Journal of Information Science and Technology* 7(1–2): 201–213.

Indyka-Piasecka, Agnieszka. (2004) *Modele użytkownika w internetowych systemach wyszukiwania informacji [User models in Internet information retrieval systems].* Ph. D. dissertation, Politechnika Wrocławska.

Israel, Glenn D. (1992) "Determining sample size". Technical report, University of Florida.

Jacquemin, Christian. (2001) *Spotting and discovering terms through Natural Language Processing.* The MIT Press.

Jain, Anil Kumar, M. Narasimha Murty and Patrick Joseph Flynn. (1999) "Data clustering: a review". *ACM Computing Surveys* 31(3): 264–323.

Joachims, Thorsten. (2002) *Learning to classify text using Supported Vector Machines.* Kluwer Academic Press.

Karypis, George. (2002) "CLUTO a clustering toolkit". Technical Report 02-017, Department of Computer Science, University of Minnesota. URL http://www.cs.umn.edu/~cluto

Kennedy, Alistair. (2006) "Analysis and construction of noun hypernym hierarchies to enhance *roget's* thesaurus". Master's thesis, School of Information Technology and Engineering, University of Ottawa.

Klavans, Judith and Min-Yen Kan. (1998) "Role of verbs in document analysis". *Proceedings of the Conference of the International Committee on Computational Linguistics*, 680–686.

Kohonen, Teuvo. (1982) "Self-organized formation of topologically correct feature maps". *Biological Cybernetics* 43(1): 59–69.

Kohonen, Teuvo, Samuel Kaski, Krista Lagus, Jarkko Salojrvi, Jukka Honkela, Vesa Paatero and Antti Saarela. (2000) "Self organization of a massive document collection". *IEEE Transactions on Neural Networks* 11: 574–585.

Kłopotek, Mieczysław A., Adam Przepiórkowski, Sławomir T. Wierzchoń and Krzysztof Trojanowski, ed. (2008) *Intelligent Information Systems XVI. Proceedings of the International IIS'08 Conference held in Zakopane, Poland, June, 2008.* Warsaw: Academic Publishing House EXIT.

Kłopotek, Mieczysław A., Sławomir T. Wierzchoń and Krzysztof Trojanowski, ed. (2006) *Intelligent Information Processing and Web Mining – Proceedings of the International IIS: IIPWM '06 Conference held in Wisła, Poland, June, 2006.* Berlin: Springer.

Kurc, Roman. (2008) “Automatyczne wydobywanie leksykalnych relacji semantycznych na podstawie prostych wzorców syntaktyczno-leksykalnych [Automatic acquisition of lexical-semantic relations using simple syntactic and lexical patterns]”. Master’s thesis, Faculty of Computer Science and Management, Wroclaw University of Technology.

Kurc, Roman and Maciej Piasecki. (2008) “Automatic Acquisition of Wordnet Relations by the Morpho-Syntactic Patterns Extracted from the Corpora in Polish”. *Proceedings of the International Multiconference on Computer Science and Information Technology – Third International Symposium Advances in Artificial Intelligence and Applications*, 181–188. URL `http://www.proceedings2008.imcsit.org/`

Landauer, Tomas K. and Susan T. Dumais. (1997) “A solution to Plato’s problem: The Latent Semantic Analysis theory of acquisition”. *Psychological Review* 104(2): 211–240.

Landwehr, Niels, Mark Hall and Eibe Frank. (2003) “Logistic Model Trees”. *Proceedings of the 14th European Conference on Machine Learning*. Croatia: Springer, 241–252.

Lapata, Maria. (2001) “A corpus-based account of regular polysemy: The case of context-sensitive adjectives”. *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*. ACL, 63–70.

Le Roux, Jurie, Koliswa Moropa, Sonja Bosch and Christiane Fellbaum. (2008) “Introducing the African languages WordNet”. Tanács et al. (2008), 269–280.

Lenci, Alessandro, Simonetta Montemagni and Vito Pirrelli. (2001) “The acquisition and representation of word meaning”. Teaching materials prepared for ESLLI’2001.

Lin, Dekang. (1993) “Principle-based parsing without overgeneration”. *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*.

———. (1998) “Automatic retrieval and clustering of similar words”. *Proceedings of the Joint Conference of the International Committee on Computational Linguistics*. ACL, 768–774.

Lin, Dekang and Patrick Pantel. (2002) “Concept discovery from text”. *Proceedings of the Joint Conference of the International Committee on Computational Linguistics*. Taipei, Taiwan, 577–583.

Lyons, John. (1989) *Semantyka [Semantics]*. PWN. Translated by A. Weinsberg

Manning, Christopher D., Prabhakar Raghavan and Hinrich Schutze. (2008) *Introduction to Information Retrieval*. Cambridge University Press. URL `http://www-csli.stanford.edu/~hinrich/information-retrieval-book.html`

Manning, Christopher D. and Hinrich Schütze. (2001) *Foundations of statistical Natural Language Processing*. The MIT Press.

Martinek, Jacek. (1997) “Pozyskiwanie informacji semantycznej ze słowników jednojęzycznych [Semantic information retrieval from unilingual dictionaries]”. Zdzisław Bubnicki and Adam Grzech, ed., *Inżynieria wiedzy i systemy ekspertowe (materiały konferencji) [Knowledge engineering and expert systems (proceedings)]*, 326–333.

Matsumoto, Yuji. (2003) *Lexical knowledge acquisition.*, Mitkov (2003)., chapter 21, 395–413.

Matsuo, Yutaka and Mitsuru Ishizuka. (2004) “Keyword extraction from a single document using word co-occurrence statistical information”. *International Journal on Artificial Intelligence Tools* 13(1): 157–169. URL `http://ymatsuo.com/papers/ijait04.pdf`

Mihalcea, Rada and Dan Moldovan. (1999) “A method for word sense disambiguation of unrestricted text”. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 152–158.

Miháltz, Márton, Csaba Hatvani, Judit Kuti, György Szarvas, János Csirik, Gábor Prószéky and Tamás Váradi. (2008) “Methods and results of the Hungarian WordNet project”. Tanács et al. (2008), 311–320. URL `http://www.inf.u-szeged.hu/projectdirs/gwc2008/`

Miller, George A. (1998) *Nouns in WordNet.*, Fellbaum (1998c)., chapter 1, 23–46.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross and Katherine J. Miller. (1990) “Introduction to WordNet: an on-line lexical database”. *International Journal of Lexicography* 3(4): 235–244. URL `ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps`

———. (1993) “Introduction to WordNet: an on-line lexical database”. Unpublished, part of the set called: “Five Papers”. URL `ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps`

Miller, George A. and Christiane Fellbaum. (2007) “WordNet then and now”. *Language Resources & Evaluation* 41: 209–214.

Miller, George A., Christiane Fellbaum, Randee Tengi, Susanne Wolff, Pamela Wakefield, Helen Langone and Benjamin Haskell. (2007) “WordNet – a lexical database for the English language”. Homepage of the project. URL `http://wordnet.princeton.edu/`

Miller, George A., Claudia Leacock, Randee Tengi and Ross T. Bunker. (1993) “A semantic concordance”. *Proceedings of the ARPA Workshop on Human Language Technology*. Princeton, USA, 303–308.

Mitchell, Tom M. (1997) *Machine Learning*. WCB McGraw-Hill.

Mitkov, Ruslan, ed. (2003) *The Oxford handbook of Computational Linguistics*. Oxford University Press.

Mohammad, Saif and Graeme Hirst. (2006) "Distributional measures as proxies for semantic relatedness". Submitted for publication in Kluwer. URL `http://ftp.cs.toronto.edu/pub/gh/Mohammad+Hirst-2005.pdf`

Mohanty, Rajat Kumar, Pushpak Bhattacharyya, Shraddha Kalele, Prabhakar Pandey, Aditya Sharma and Mitesh Kopra. (2008) "Synset based multilingual dictionary: Insights, applications and challenges". Tanács et al. (2008), 321–333.

Moldovan, Dan I. and Rada Mihalcea. (2000) "Using WordNet and lexical operators to improve internet searches". *IEEE Internet Computing* 4(1): 34–43. URL `http://www.seas.smu.edu/~rada/papers/int-comp.99.ps.gz`

Morato, Jorge, Miguel Ángel Marzal, Juan Lloréns and José Moreiro. (2004) "WordNet Applications". *Proceedings of the Second Global WordNet Conference*, 270–278. URL `http://www.fi.muni.cz/gwc2004/proc/105.pdf`

Morin, Emmanuel and Christian Jacquemin. (1999) "Projecting corpus-based semantic links on a thesaurus". *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 389–396.

———. (2004) "Automatic acquisition and expansion of hypernym links". *Computer and the Humanities* 38(4): 343–362.

Niles, Ian and Adam Pease. (2001) "Towards a standard upper ontology". Chris Welty and Barry Smith, ed., *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems*. Ogunquit, Maine, 2–9.

Obrębski, Tomasz. (2002) *Automatyczna analiza składniowa języka polskiego z wykorzystaniem gramatyki zależnościowej [Automatic syntactic analysis of Polish using a dependency grammar]*. Ph. D. dissertation, Instytut Podstaw Informatyki PAN.

Pala, Karel and Pavel Smrž. (2004) "Building Czech Wordnet". *Romanian Journal of Information Science and Technology* 7(1–2): 79–88.

Palmer, Martha, Daniel Gildea and Paul Kingsbury. (2005) "The proposition bank: A corpus annotated with semantic roles". *Computational Linguistics* 31(1): 71–106.

Pantel, Patrick. (2003) *Clustering by committee*. Ph. D. dissertation, Department of Computing Science, University of Alberta. Adviser-Dekang Lin.

Pantel, Patrick and Marco Pennacchiotti. (2006) "Esspresso: Leveraging generic patterns for automatically harvesting semantic relations". Calzolari et al. (2006), 113–120.

Pantel, Patrick and Deepak Ravichandran. (2004) "Automatically labeling semantic classes". Daniel Marcu Susan Dumais and Salim Roukos, ed., *Proceedings of the Human Language Technology Conference / Annual Meeting of the North American Chapter of the Association for Computational Linguistics*. Boston, Massachusetts, USA: Association for Computational Linguistics, 321–328.

Pantel, Patrick, Deepak Ravichandran and Eduard Hovy. (2004) “Towards terascale knowledge acquisition”. *Proceedings of the Joint Conference of the International Committee on Computational Linguistics*. Geneva, Switzerland, 771–777. URL `http://www.patrickpantel.com/cgi-bin/Web/Tools/getfile.pl?type=paper&id=2004/coling04.pdf`

Pedersen, Bolette Sandford and Sanni Nimb. (2008) “Event hierarchies in DanNet”. Tanács et al. (2008), 339–348.

Pedersen, Ted. (2006) “Unsupervised corpus based methods for WSD”. Agirre and Edmonds (2006), 133–166.

Piasecki, Maciej. (2006) “Handmade and automatic rules for Polish tagger”. Sojka et al. (2006), 205–212.

Piasecki, Maciej and Bartosz Broda. (2007) “Semantic similarity measure of polish nouns based on linguistic features”. Witold Abramowicz, ed., *Proceedings of the 10th International Conference on Business Information Systems*. Springer, 381–390.

Piasecki, Maciej and Grzegorz Godlewski. (2006) “Effective architecture of the Polish tagger”. Sojka et al. (2006), 213–220.

Piasecki, Maciej and Paweł Koczan. (2007) “Environment supporting construction of the Polish WordNet”. Vetulani (2007), 519–523.

Piasecki, Maciej, Michał Marcińczuk, Stanisław Szpakowicz and Bartosz Broda. (2008) “Classification-based filtering of semantic relatedness in hypernymy extraction”. *Proceedings of the Sixth International Conference on Natural Language Processing – GoTAL 2008*. Springer, 393–404.

Piasecki, Maciej and Adam Radziszewski. (2009) “Morphosyntactic constraints in acquisition of linguistic knowledge for Polish”. Agnieszka Mykowiecka and Małgorzata Marciniak, ed., *Aspects of Natural Language Processing*. Springer, 163–190. *Accepted, volume in preparation*. Essays dedicated to Professor Leonard Bolc.

Piasecki, Maciej, Stanisław Szpakowicz and Bartosz Broda. (2007a) “Automatic selection of heterogeneous syntactic features in semantic similarity of Polish nouns”. *Proceedings of the Text, Speech and Dialog Conference*. Springer, 66–75.

———. (2007b) “Extended similarity test for the evaluation of semantic similarity functions”. Vetulani (2007), 104–108.

Piotrowski, Tadeusz and Zygmunt Saloni. (1999) *Kieszonkowy słownik angielsko-polski i polsko-angielski [Pocket English-Polish and Polish-English Dictionary]*. Warszawa: Wyd. Wilga.

Przepiórkowski, Adam. (2004) *The IPI PAN Corpus, preliminary version*. Institute of Computer Science PAS.

———. (2006) “The potential of the IPI PAN Corpus”. *Poznań Studies in Contemporary Linguistics* 41: 31–48.

Pucher, Michael. (2007) “WordNet-based semantic relatedness measures in automatic speech recognition for meetings”. *Proccedings of the 45th Annual Meeting of the Association for Computational Linguistics*. Prague, Czech Republic: ACL, 129–132.

PWN. (2007) “Słownik języka polskiego”. Published on the web page. URL `http://sjp.pwn.pl/`

Quinlan, Ross. (1986) “Induction of decision trees”. *Machine Learning* 1(1): 81–106.

Raffaelli, Ida, Marko Tadić, Božo Bekavac, and Željko Agić. (2008) “Building Croatian WordNet”. Tanács et al. (2008), 349–359.

Rauber, Andreas, Dieter Merkl and Michael Dittenbach. (2002) “The growing hierarchical self-organizing maps: exploratory analysis of high-dimensional data”.

Ravichandran, Deepak and Eduard Hovy. (2002) “Learning surface text patterns for a question answering system”. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, PA: ACL, 41–47.

Ravichandran, Deepak, Patrick Pantel and Eduard Hovy. (2002) “Randomized algorithms and NLP: Using locality sensitive hash function for high speed noun clustering”. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 622–629.

Rodríguez, Horacio, David Farwell, Javi Farreres, Manuel Bertran, Musa Alkhalifa, M. Antonia Martí, William Black, Sabri Elkateb, James Kirk, Adam Pease, Piek Vossen and Christiane Fellbaum. (2008) “Arabic WordNet: Current state and future extensions”. Tanács et al. (2008), 387–405.

Rosenzweig, Joseph, Rada Mihalcea and Andras Csomai. (2007) “WordNet bibliography”. Web page: a bibliography referring to research involving the WordNet lexical database. URL `http://lit.csci.unt.edu/%7Ewordnet`

Rubenstein, Herbert and John B. Goodenough. (1965) “Contextual correlates of synonymy”. *Communication of the ACM* 8(10): 627–633.

Ruge, Gerda. (1992) “Experiments on linguistically-based term associations”. *Information Processing and Management* 28(3): 317–332.

Ruppenhofer, Josef, Collin F. Baker and Charles J. Fillmore. (2002) “The FrameNet database and software tools”. Anna Braasch and Claus Povlsen, ed., *Proceedings of the Tenth Euralex International Congress*. Copenhagen: Denmark, 371–375.

Ryu, Pum-Mo and Key-Sun Choi. (2006) “Taxonomy learning using term specificity and similarity”. *Proceedings of the Second Workshop on Ontology Learning and Population*. Sydney: ACL, 41–48.

Rzeczpospolita (2008) “Korpus Rzeczpospolitej”. [on-line] `www.cs.put.poznan.pl/dweiss/rzeczpospolita` Corpus of text from the online edtion of Rzeczpospolita.

Sahlgren, Magnus. (2001) “Vector-based semantic analysis: Representing word meanings based on random labels”. *Proceedings of the Semantic Knowledge Acquisition and Categorisation Workshop, ESSLLI 2001*. Helsinki, Finland.

Schütze, Hinrich. (1998) “Automatic word sense discrimination”. *Computational Linguistics* 24(1): 97–123.

Shamsfard, Mehrnoush. (2008) “Developing FarsNet: A lexical ontology for Persian”. Tanács et al. (2008), 413–418.

Sinha, Manish, Mahesh Reddy and Pushpak Bhattacharyya. (2006) “An approach towards construction and application of multilingual Indo-WordNet”. *Proceedings of the Third Global WordNet Conference*, 259–263.

Snow, Rion, Daniel Jurafsky and Andrew Y. Ng. (2005) “Learning syntactic patterns for automatic hypernym discovery”. Lawrence K. Saul *et al.*, ed., *Advances in Neural Information Processing Systems 17*. Cambridge, MA: MIT Press, 1297–1304. URL `http://www.stanford.edu/~jurafsky/paper887.pdf`

Snow, Rion, Dan Jurafsky and Andrew Y. Ng. (2006) “Semantic taxonomy induction from heterogenous evidence”. Calzolari et al. (2006), 801–808.

Sojka, Petr, Ivan Kopecek and Karel Pala, ed. (2006) *Proceedings of the Text, Speech and Dialog 2006 Conference*. Springer.

Sojka, Petr, Karel Pala, Pavel Smrž, Christiane Fellbaum and Piek Vossen, ed. (2004) *Proceedings of the Second International WordNet Conference*. Masaryk University Brno, Czech Republic.

Strapparava, Carlo and Alessandro Valitutti. (2004) “WordNet-Affect: an affective extension of WordNet”. *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, 1083–1086.

Tanács, Attila, Dóra Csendes, Veronika Vincze, Christiane Fellbaum and Piek Vossen, ed. (2008) *Proceedings of the Fourth Global WordNet Conference*. University of Szeged, Department of Informatics.

Tengi, Randee I. (1998) *Design and implementation of the WordNet lexical database and searching software.*, Fellbaum (1998c)., chapter 4, 105–127.

Tomuro, Noriko, Steven L. Lytinen and Hitoshi Isahara Kyoko Kanzaki. (2007) “Clustering using feature domain similarity to discover word senses for adjectives”. *Proceedings of the First IEEE International Conference on Semantic Computing*. IEEE, 370–377.

Tufiş, Dan, Dan Cristea and Sofia Stamou. (2004) “BalkaNet: Aims, methods, results and perspectives. a general overview”. *Romanian Journal of Information Science* *and Technology* 7(1–2): 9–43 Special Issue. URL `http://www.ceid.upatras.gr/Balkanet/journal/7_Overview.pdf`

Turney, Peter D. (2001) “Mining the Web for synonyms: PMI-IR versus LSA on TOEFL”. *Proceedings of the Twelfth European Conference on Machine Learning*. Berlin: Springer-Verlag, 491–502.

Turney, Peter D., Michael L. Littman, Jeffrey Bigham and Victor Shnayder. (2003) “Combining independent modules to solve multiple-choice synonym and analogy problems”. *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. Borovets, Bulgaria, 482–489.

Veale, Tony and Yanfen Hao. (2008) “Enriching WordNet with folk knowledge and stereotypes”. Tanács et al. (2008), 453–461.

Vetulani, Zygmunt, ed. (2007) *Proceedings of the Third Language and Technology Conference*. Wydawnictwo Poznańskie Sp. z o.o.

Vetulani, Zygmunt, Justyna Walkowska, Tomasz Obrębski, Paweł Konieczka, Przemysław Rzepecki and Jacek Marciniak. (2007) “PolNet – Polish WordNet project algorithm”. Vetulani (2007), 172–176.

Vossen, Piek. (2002) “EuroWordNet general document version 3”. Raport, University of Amsterdam.

———. (2003) *Ontologies.*, Mitkov (2003)., chapter 25, 464–482.

Vossen, Piek, Eneko Agirre, Nicoletta Calzolari, Christiane Fellbaum, Shu-Kai Hsieh, Chu-Ren Huang, Hitoshi Isahara, Kyoko Kanzaki, Andrea Marchetti, Monica Monachini, Federico Neri, Remo Raffaelli, German Rigau, Maurizio Tesconi and Joop VanGent. (2008) “KYOTO: A system for mining, structuring, and distributing knowledge across languages and cultures”. Tanács et al. (2008), 474–484.

Vossen, Piek, 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.

Weeds, Julie and David Weir. (2005) “Co-occurrence retrieval: A flexible framework for lexical distributional similarity”. *Computational Linguistics* 31(4): 439–475.

Widdows, Dominic. (2003) “Unsupervised methods for developing taxonomies by combining syntactic and statistical information”. *Proceedings of the Human Language Technology Conference / Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, 197–204.

———. (2004) *Geometry and meaning*. CSLI Publications.

Wierzbicka, Anna. (2000) *Język–umysł–kultura [Language–mind–culture]*. PWN.

Witschel, Hans Friedrich. (2005) "Using decision trees and text mining techniques for extending taxonomies". *Proceedings of the Workshop on Learning and Extending Lexical Ontologies by using Machine Learning Methods at ICML-05*, 61–68. URL http://wortschatz.uni-leipzig.de/~fwitschel/papers/WS_ontologies.pdf

Witten, Ian H. and Eibe Frank. (2005) *Data mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann, 2nd ed.

Woliński, Marcin. (2005) "An efficient implementation of a large grammar of Polish". *Archives of Control Sciences* 15(LI)(3): 251–258.

———. (2006) "Morfeusz – a practical tool for the morphological analysis of Polish". Kłopotek et al. (2006), 511–520.

Zaiane, Osmar R., Eli Hagen and Jiawei Han. (1999) "Word taxonomy for online visual asset management and mining". *Proceeding of the Fourth International Conference on Applications of Natural Language to Information Systems – NLDB'99*, 271–275.

Zesch, Torsten and Iryna Gurevych. (2006) "Automatically creating datasets for measures of semantic relatedness". *Proceedings of the Workshop on Linguistic Distances*. Sydney, Australia: Association for Computational Linguistics, 16–24.

Zhang, Min, Jie Zhang and Jian Su. (2006) "Exploring syntactic features for relation extraction using a convolution tree kernel". *Proceedings of the Human Language Technology Conference / Annual Meeting of the North American Chapter of the Association for Computational Linguistics*. ACL, 288–295.

Zhuang, Li and Xiaoyan Zhu. (2005) "An OCR post-processing approach based on multi-knowledge". *Knowledge-Based Intelligent Information and Engineering Systems*, volume 3681 series *LNCS*. Berlin / Heidelberg: Springer, 346–352.

List of Tables

List of Figures

Index