Lexico-Semantic Annotation of *Składnica* Treebank by means of PLWN Lexical Units

Elżbieta Hajnicz
Institute of Computer Science, Polish Academy of Sciences
ul. Ordona 21, 01-237 Warsaw, Poland
hajnicz@ipipan.waw.pl

Abstract

In this paper we present the principles of lexico-semantic annotation of *Składnica* Treebank using Polish WordNet lexical units. We describe different means of annotation, depending on the structure of a sentence in *Składnica* on the one hand and the availability of adequate lexical unit in PLWN on the other. Apart from "standard" annotation involving lexical units with the same lemma as the token under annotation, multi-word units, different verb lemmas including reflexive marker *się* as well as synonyms and hypernyms have also been involved. Some tokens have obtained tags explaining why they require no annotation. Additionally, we discuss the assessment of the annotation of whole sentences.

1 Introduction

It is widely acknowledged that linguistically annotated corpora play a crucial role in NLP. There is even a tendency towards their ever-deeper annotation. In particular, semantically annotated corpora become more and more popular, because they have several applications in word sense disambiguation (Agirre and Edmonds, 2006) or automatic construction of lexical resources (McCarthy, 2001; Schulte im Walde, 2006; Sirkayon and Kawtrakul, 2007). The important part of semantically annotated corpora are semantically annotated treebanks.

In this paper, the procedure of lexico-semantic annotation of *Składnica* Treebank (cf. section 3.1), the largest Polish treebank, is presented. Verbal, nominal and adjectival tokens forming sentences are annotated using Polish WordNet (PLWN, cf. section 3.2) lexical units. Special attention is paid to tokens for which a correct interpretation was not found in the wordnet.

The annotation is performed using a dedicated tool *Semantikon* (Hajnicz, 2013c). Each sentence is annotated by two linguists, and conflicts are resolved by a master linguist.

The procedure of lexico-semantic annotation of *Składnica* was preceded by tagging named entities with corresponding PLWN-base semantic types (Hajnicz, 2013b), by means of semi-automatic transfer of information from the NE annotation layer (Savary et al., 2010) of the National Corpus of Polish (NKJP). Unlike with common words, this information was linked to nonterminal nodes, since named entities are very often multi-word units. For NEs present in PLWN, corresponding lexical units were used, other NEs were tagged by means of synset identifiers corresponding to their semantic types.

Section 2 presents related work on semantic annotation of text corpora. Section 3 contains the description of resources used. The principles of the actual annotation of tokens are discussed in section 4, whereas the rules of the assessment of whole sentences are presented in section 5.

2 Semantically annotated corpora

Semantic annotation of text corpora seems to be the last phase in the process of corpus annotation, less popular than morphosyntactic and (shallow or deep) syntactic annotation. However, there exist semantically annotated subcorpora for many languages, some of them wordnet-based. They are usually substantially smaller than other types of corpora.

The most famous semantically annotated corpus is SemCor (Miller et al., 1993). It is a subcorpus of the Brown Corpus (Francis and Kucera, 1964) containing 250 000 words semantically annotated using Princeton WordNet (PWN) (Miller et al., 1990; Fellbaum, 1998; Miller and Fellbaum, 2007, http://wordnet.

princeton.edu/) synset identifiers. The annotation includes proper names and collocations (the ones present in PWN). A special tag is assigned for tokens with no available sense considered appropriate (supplemented with a corresponding comment).

For Polish, lexico-semantic annotation was performed for the sake of experiments in WSD, and was limited to small sets of highly polysemic words (Broda et al., 2009; Kobyliński, 2011; Przepiórkowski et al., 2011), first of them using PLWN lexical units.

Unlike other corpora, semantic annotation of treebanks usually are not limited to lexico-semantic annotation. Nevertheless, there exist some lexico-semantically annotated treebanks. In particular, a fragment of the Penn Treebank was lexico-semantically tagged by means of PWN senses (Palmer et al., 2000). The Portuguese Treebank *Floresta sintá(c)tica* (Alfonso et al., 2002) was annotated by means of a predefined hierarchy of semantic tags called *semantic prototypes* (Bick, 2006).

An interesting example is the Italian Syntactic-Semantic Treebank (Montemagni et al., 2003b; Montemagni et al., 2003a), which lexico-semantic annotation is based on ItalWordNet (IWN) (Roventini et al., 2000) sense repository being a part of EuroWordNet. When more than one IWN sense applies to the context being tagged, underspecification is allowed (expressed by disjunction/conjunction of senses). Special tags allow marking the lack of a corresponding sense in IWN, metaphoric or methonymic usage of words or expressions, diminutive and augmentative derivatives, and idioms. Moreover, named entities are tagged with their (rather coarse) semantic types.

3 Data resources

Presented work is based on two resources: the Polish Treebank *Składnica* and the Polish Wordnet called *Słowosieć* (English acronym PLWN).

3.1 *Składnica*

Składnica (Świdziński and Woliński, 2010; Woliński et al., 2011) is a bank of constituency parse trees for Polish sentences taken from selected paragraphs in the balanced manually-annotated subcorpus of the Polish National Corpus (NKJP). To attain consistency of the treebank, a semi-automatic method was applied: trees were generated by an automatic parser[1] and then selected and validated by human annotators. The resulting version 0.5 of *Składnica* contains 8241 manually validated trees.

As a consequence of the method used, some sentences do not have any correct parse tree assigned, if *Świgra* did not generate any tree for a particular sentence or no generated tree has been accepted as correct one.

Parse trees are encoded in XML, each parse being stored in a separate file. The parse tree of sentence *Taki był u nas zwyczaj od pokoleń.* ('There was such a habit among us for generations.') in *Składnica* is shown in Fig. 1.

3.2 Polish wordnet—*Słowosieć*

In contrast to NKJP annotation, we decided to annotate tokens with very fine-grained semantic types represented by wordnet synsets. For this goal, we used PLWN (Piasecki et al., 2009).

PLWN is a network of lexico-semantic relations, an electronic thesaurus with a structure modelled on that of the Princeton WordNet and those constructed in the EuroWordNet project. Polish WordNet describes the meaning of a lexical unit comprising one or more words by placing this unit in a network representing relations such as synonymy, hypernymy, meronymy, etc.

A lexical unit (LU) is a string which has its morphosyntactic characteristics and a meaning as a whole. Therefore, it may be an idiom or even a collocation, but not a productive syntactic structure (Derwojedowa et al., 2008). An LU is represented as a pair ⟨lemma, meaning⟩, the last being a natural number. Technically, any LU also has its unique numeric identifier. Each lexical unit belongs to a synset, which is a set of synonyms. Synsets have their unique numeric identifiers as well. A fragment of the table of triples ⟨identifier, lemma, meaning⟩ is presented in Fig. 2.

Version 2.0 of PLWN is used for the semantic annotation of tokens. It contains 106438 lemmas, namely 17486 verb lemmas, 77662 noun lemmas and 11290 adjective lemmas, 32199 of them (7234 verb, 20625 noun and 4340 adjective lemmas) being ambiguous. The number of lexical units is 160100 (31980 verb, 109967 noun and 18153 adjective units). On the other hand, named entity annotation was performed by means of PLWN 1.6.

[1] *Świgra* parser (Woliński, 2005) based on the revised version (Świdziński and Woliński, 2009) of metamorphosis grammar GFJP (Świdziński, 1992).

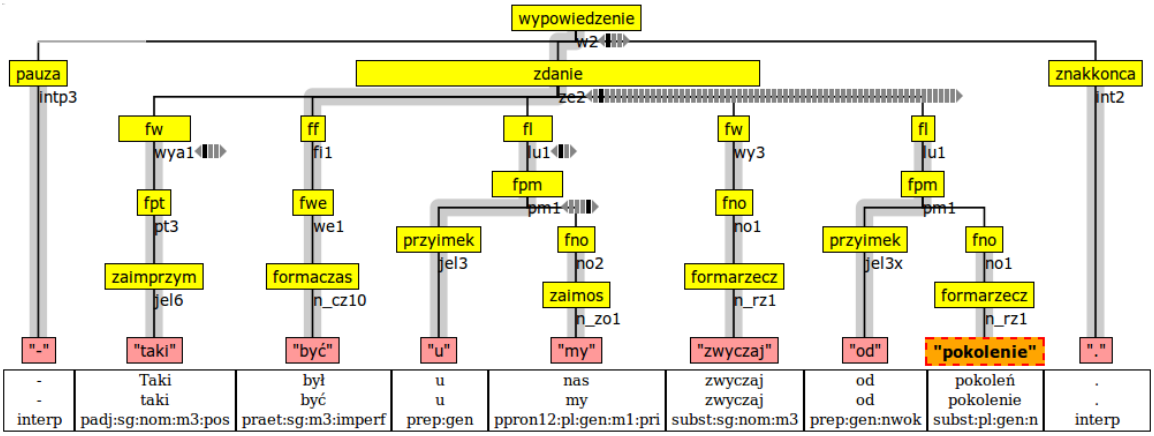

Figure 1: Exemplary parse tree from *Składnica*

```
124     aparycja        1
136     apteka          1
139     arbiter         2
198     atrybut         3
199     atrybut         1
18382   atrybut         2
19474   arbiter         1
```

Figure 2: The fragment of the table of triples ⟨identifier, lemma, meaning⟩ of PLWN 1.6

3.2.1 Named entities in PLWN

Polish WordNet contains some number of named entities, selected rather randomly. They are represented in the same way as common words, by means of lexical units. LUs representing NEs are grouped in synsets as well, since the same object can be identified by means of several NEs (e.g., a full name and its acronym). The only difference is that they are connected by 'type' and 'instance' relations instead of 'hypernym' and 'hyponym'.

The representation of NEs in PLWN is far from satisfactory. Therefore, a table of names (a sort of a gazetteer) has been created, in which a list of semantic types represented by PLWN synset identifiers is assigned to every NE lemma. The order of synsets in a list reflects their preference.

4 Principles of annotation

4.1 The scope of annotation

PLWN contains lexical units of three open parts of speech: adjectives, nouns and verbs. Therefore, only tokens belonging to these POS are annotated. This concerns abbreviations and acronyms as well[2].

[2]Acronyms usually are named entities.

Unfortunately, it does not contain adverbs so far, hence we have no possibility of annotating them. This causes a kind of inconsistency in annotation, which we hope to correct in the future.

On the other hand, only sentences having parse trees are annotated. The reason for this is that corresponding LUs are assigned to terminal nodes representing tokens being annotated. This feature can limit applicability of the resulting resource in WSD.

In the case of tokens being elements of multiwords named entities, the human annotators were free to decide whether they should be annotated. The reason is that some NEs (mainly names of institutions) are compositional.

Semantic annotation is introduced into XML structure of a parse tree as a new type child element of the element `node`: a terminal node (element `plwn_interpretation`) for common words and a nonterminal node (element `named`) for named entities. All LUs from PLWN with the corresponding lemma (and POS) are included, the correct ones having the attribute `chosen="true"` (see Fig. 3 for the noun *pokolenie—generation*). The attribute `polysemy` is used to indicate whether the list of lemmas is a singleton or not. Storing all LUs enables to check what choices were accessible for human or automatic annotators during the process of annotation. The actual annotation is not ambiguous.

In PLWN, there are also units whose lemmas differ only in letter case (lower- vs. uppercase). If the attribute `case_agreement` has the value `true`, only LUs with the lemma identical with the token lemma are considered. Otherwise, the chosen LU lemma differs from the token lemma

```
<plwn_interpretation sem_id="sem_5">
    <plwn_units case_agreement="true"
                polysemy="true">
        <unit luid="sem_5-sv1"
              chosen="true">
            <lubase>pokolenie</lubase>
            <lusense>1</lusense>
            <luident>20791</luident>
            <synset>2418</synset>
        </unit>
        <unit luid="sem_5-sv2">
            <lubase>pokolenie</lubase>
            <lusense>2</lusense>
            <luident>5921</luident>
            <synset>7789</synset>
        </unit>
    </plwn_units>
</plwn_interpretation>
```

Figure 3: XML representation of a polysemic common word

in that aspect (and all corresponding LUs are included).

Additionally, the root element is augmented with three attributes, `name-plwn_version`, `sense-plwn_version`, `final-plwn_version` pointing out which version of PLWN was used for a particular phase of semantic annotation. Certainly, it is possible that these three parameters are equal, but since both resources are under long-lasting intensive manual development, this is highly unlikely. The procedure of updating the annotation to the current version of resources (Hajnicz, 2013a) has been elaborated (the third attribute).

The Table 1 summarises the XML elements ant their attributes used for lexico-semantic level of annotation. The element `plwn_units` is used for standard annotation, as in Fig. 3, the element `other_units` is used for synonyms, hypernyms, multi-element units etc., whereas the element `derived_units` is used for gerunds and participles (see Fig. 4). The attributes `type`, `relat`, and `chosen` are optional; the attributes `deriv_type` and `deriv_dest` appear in `plwn_units` only if the element `derived_units` is present (see section 4.2.4).

4.2 Non-standard annotation

Apart from the standard annotation involving lexical units of the same lemma as a token itself, some tokens are tagged in a special way, including:

- multi-word units,
- verb lemmas including reflexive marker *się*,
- synonyms and hypernyms.

For such annotations, the XML element `other_units` instead of `plwn_units` is used, having the attribute `relat` determining the type of special annotation.

If LUs having the same lemma as a token under annotation occur in PLWN, then the corresponding `plwn_units` element appears in the corresponding `plwn_interpretation`. However, no of its units are provided with the attribute `chosen="true"`, as they were not adequate interpretation of a token in a particular context. Note also that the attribute `case_agreement` is not considered for`other_units`, as the lemma of LUs is different from the lemma of a token, hence their case cannot be compared.

4.2.1 Multi-word units

PLWN contains a growing number of multi-word units. In PLWN 2.0, 12% of units have multi-word lemmas: (15% nouns LUs, 5% verb LUs and only 0.2% adjective LUs). There are two kinds of such units:

- units specifying the meaning of the head of lemma, e.g., *szkoła podstawowa* ('primary school') is a school; such LUs are hyponyms of units representing the head of their lemmas;
- units changing the meaning of the head of lemma, e.g., *centrum handlowe* ('shopping centre') is not a *centre*; such LUs are not connected with any unit representing the head of their lemmas.

In the first case, the annotation of tokens using the single-word hypernym is correct, even though less precise. In the second case, using a multi-word expression is indispensable to obtain the correct annotation. In any case, the attribute `relat` gets the value `multi-unit`.

As in the standard case, multi-word LU annotation is attached to individual tokens. The reason for this is twofold. First, due to its structure, *Składnica* may not contain a single node corresponding to the relevant expression. For instance, the expression *szkoły podstawowej w Tychnowach* ('primary school in Tychnowy') from the sentence *Adam [...] chodzi do III klasy szkoły podstawowej w Tychnowach* ('Adam attends the III class of the primary school in Tychnowy'), is represented in *Składnica* by a single node, having three child

Table 1: XML representation for lexico-semantic level of annotation

elements	attribute	values
plwn_interpretation	sem_id	`identifier`
	type	multi-element, grammatical, foreign, lack, neologism, prep-element, wrong-lemma
plwn_units, derived_units, other_units	polysemy	true, false
plwn_units, derived_units	case_agreement	true, false
	deriv_type	ger, pact, ppas
plwn_units	deriv_dest	`lemma`
derived_units	deriv_source	`lemma`
other_units	relat	refl, multi-unit, synonym, hypernym
unit	luid	`identifier`
	chosen	true, match

nodes corresponding to *szkoły* ('school'), *podstawowej* ('primary') and *w Tychnowach* ('in Tychnowy'), and no node corresponding to *szkoły podstawowej* ('primary school'). Secondly, there are sentences in which only the heads of such expressions occur (e.g., *Lubimy zaglądać do takich dużych centrów*—'We like to visit such big [shopping] centres').

If a multi-word expression (present in PLWN) is semantically compositional, its non-head elements are annotated in the standard way. Otherwise, the element `plwn_interpretation` obtains the attribute `type="multi-element"`.

4.2.2 Verb lemmas with the reflexive marker

As in other Slavic languages, in Polish, the reflexive marker *się* can form an integral part of the lemma of a verb[3]. In Polish, *się* is a separate orthographic word, not attached to a verb. Verbs with and without *się* included in their lemma have different meaning and are represented by means of separate LUs. For instance, *zalecać* means 'to recommend, to order', whereas *zalecać się* means 'to make advances (to somebody)'. 9% of LUs have lemmas with the reflexive marker (23% of verbs, 6,5% of nouns: 23% of gerunds, as could be expected).

If a verb token is annotated in such a way, its annotation contains the attribute `relat="refl"`. It is considered separately from typical multi-word expressions, since it is a linguistic feature completely different and independent from collocations. In particular, there are verbal multi-word expressions in spite of the occurrence of the reflexive marker (e.g., *podać się do dymisji*—'to demit').

[3]Some occurrences of *się*, namely impersonal, strictly reflexive and reciprocal, are not part of a verb lemma.

4.2.3 Synonyms and hypernyms

It is almost impossible that there is a corresponding lexical unit in PLWN for every token in *Składnica*, since both words and their meanings exhibit Zipfian distribution, the more so as PLWN is a resource under intensive development.

SemCorr and the Italian Syntactic-Semantic Treebank apply special tags for such tokens. However, such a solution limits the information about the missing senses to informal textual comments. We decided to introduce annotation using synonyms or hypernyms. Such annotation locates the absent meaning of a word in a structure of PLWN as precisely as possible. The attribute `relat` of the corresponding `other_units` element gets the value `synonym` or `hypernym`, respectively.

Hypernyms are used if synonyms of absent LUs do not occur in PLWN. Usually, synonyms for absent noun units are proportionally easy to establish, but adjective units and verb units are approximated by their hypernyms much more often.

The annotation by means of synonyms and hypernyms is used for tokens lemmatised improperly in *Składnica* (`type="wrong-lemma"`), and for foreign-language words tagged morphosyntactically as verbs, nouns or adjectives (`type="foreign"`).

This kind of annotation allows for finding a correct interpretation of tokens by means of newly-added LUs during an update of lexico-semantic annotation of *Składnica* to the new version of PLWN (Hajnicz, 2013a).

A similar procedure is applied for spelling errors (`type="spelling"`). The difference between spelling errors and improper lemmatisations is that the latter are supposed to be corrected.

4.2.4 Gerunds and participles

Gerunds and participles are lemmatised to verb lemmas in *Składnica*, hence they have obtained a verb interpretation. Nevertheless, they occur in sentences in nominal and adjectival positions, hence it would be natural to interpret them as nouns and adjectives, respectively.

PLWN 2.0 contains a lot of gerunds (27% of noun units) and a considerably smaller amount of participles (1.2% of adjective units). Each of them is connected with the verb unit it is derived from by means of inter-paradigmatic synonymy. Therefore, they obtain double interpretation, both by means of verbal and nominal/adjectival units (see Fig. 4 for the gerund *funkcjonowanie—functioning*).

4.3 Tokens without semantic interpretation

The procedure of annotation assumes providing as many verb, noun and adjective tokens with lexico-semantic annotation as possible. However, there are some exceptions to this rule. First, individual elements of named entities and multiwords expression need not be interpreted, having the attribute `type` equal to `name-element` or `multi-element`, respectively. For the tokens for which finding an interpretation (even by means of a hypernym) fails, this attribute equals `lack`.

Next, tokens having a grammatical function in a sentence only are not semantically interpreted and tagged as `grammatical`. This concerns mainly future forms of the verb *być (to be)* forming future tense, e.g., *Zarobki wszystkich nauczycieli* ***będą*** *rosły co rok* ('Earnings of all teachers will grow every year'), forms of the verb *być* ('to be; will') and *zostać* ('to become') forming passive voice, e.g., *Maciej R. został już dyscyplinarnie zwolniony* ('Maciej R. was already dismissed on grounds of discipline'). Non-anaphoric occurrences of pronouns are treated in the same way.

In Polish, there exist compound prepositions composed of a simple pronoun and a noun, e.g., *na temat* ('on the subject of'). Some of them were represented in *Składnica* as standard PPs, with their NP complement represented as a modifier of the noun element of the whole preposition. Such mistagged tokens have not been not semantically interpreted, obtaining instead the attribute `type="prep-element"`.

5 Assessment of a sentence

In spite of lexico-semantic interpretation at the level of single tokens, the assessment procedure involves annotation of a whole sentence. There are following assessment marks:

1. fully annotated sentence,
2. lack of corresponding lemma,
3. lack of corresponding LU,
4. occurrence of anaphora,
5. occurrence of ellipsis,
6. occurrence of metaphor,
7. occurrence of metonymy,
8. incorrect lemmatisation of a token,
9. incorrect sentence.

The first category requires that the annotation of all autosemantic tokens in the sentence is correct and final, the last one means that the sentence has not been annotated at all. Other marks concern particular problems and phenomena occurring in the sentence, hence several such marks can be attached to it, forming a list of assessments. In particular, the 3rd assessment means that there is no lexical unit in PLWN corresponding to a particular word meaning in context, whereas the 2nd assessment means that the whole lemma was not considered in PLWN.

We decided to attach information about metaphorical or metonymical usage to whole sentences instead of tokens, contrary to the Italian Syntactic-Semantic Treebank. The reason for this is that, in our opinion, they are expressed through the relations between the words rather than through any particular words.

The assessments can be used for several purposes. First, the user can search *Składnica* for sentences having particular features (i.e, metaphorical ones). Second, the information of lacking LUs and whole lemmas can be used for PLWN development and updating *Składnica* to new versions of PLWN (Hajnicz, 2013a). Finally, such an information can be used for WSD training and evaluating, and for determining selectional preferences of predicates, we are particularly interested in.

```
<plwn_interpretation sem_id="sem_2">
    <plwn_units case_agreement="true" polysemy="false"
                deriv_type="ger" deriv_dest="funkcjonowanie">
        <unit luid="sem_2-sv1" chosen="match">
            <lubase>funkcjonować</lubase>
            <lusense>1</lusense>
            <luident>1824</luident>
            <synset>54227</synset>
        </unit>
    </plwn_units>
    <derived_units case_agreement="true" polysemy="false"
                   deriv_type="ger" deriv_source="funkcjonować">
         <unit luid="der_2-sv1" chosen="true">
             <lubase>funkcjonowanie</lubase>
             <lusense>1</lusense>
             <luident>126208</luident>
             <synset>91200</synset>
        </unit>
    </derived_units>
</plwn_interpretation>
```

Figure 4: XML representation of a gerund semantic interpretation

6 Conclusions

In this paper, we have presented the principles of lexico-semantic annotation of *Składnica* Treebank by means of Polish WordNet lexical units. We have devoted the most attention to issues connected with PLWN usage.

The procedure of semantic annotation of *Składnica* is not finished yet. The 8283 sentences in *Składnica* contains 49264 nouns, verbs and adjectives for annotation, and 17410 of them belonging to 2785 (34%) sentences has been already annotated. For 2072 tokens (12%), the LU appropriate in the context has not been found in PLWN.

Applying annotation by means of (potential) synonyms or hypernyms of units absent in PLWN seems to be the main novelty of our approach, the more so as PLWN is a resource still under intensive development. Therefore, sentence assessments allow for easily finding the set of sentences containing tokens without a final interpretation, whereas synonyms and hypernyms used for their approximate annotation will facilitate their localisation in the PLWN structure.

PLWN contains a rich set of lexical and synset relations, including diminutive, augmentative, feminine derivatives, etc. Such relations could be used in the case of absence of the LU appropriate for a token, in spite of synonyms and hypernyms. However, this would further complicate the process of annotation and, as a consequence, increase the risk of errors during manual annotation. Similarly, we resigned from using interparadigmatic synonymy and hypernymy for annotating derivatives belonging to different POS.

More details about the procedure and the results of manual annotation could be found in (Hajnicz, 2013c).

Acknowledgements This research is supported by the POIG.01.01.02-14-013/09 project which is co-financed by the European Union under the European Regional Development Fund.

References

Eneko Agirre and Philip Edmonds, editors. 2006. *Word Sense Disambiguation. Algorithms and Applications*, volume 33 of *Text, Speech and Language Technology*. Springer-Verlag, Dordrecht, the Netherlands.

Susana Alfonso, Eckhard Bick, Renato Haber, and Diana Santos. 2002. Floresta sintá(c)tica: a treebank of portuguese. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, pages 1698–1703, Las Palmas, Spain.

Eckhard Bick. 2006. Noun sense tagging: Semantic prototype annotation of a portuguese treebank. In Jan Hajič and Joakim Nivre, editors, *Proceedings of the 5th Workshop on Treebanks and Linguistic Theories*, pages 127–138, Prague, Czech Republic.

Bartosz Broda, Maciej Piasecki, and Marek Maziarz. 2009. Evaluating LexCSD—a weakly-supervised method on improved semantically annotated corpus in a large scale experiment. In Mieczysław A. Kłopotek, Małgorzata Marciniak, Agnieszka Mykowiecka, Wojciech Penczek, and Sławomir T. Wierzchoń, editors, *Intelligent Information Systems*, Challenging Problems in Science: Computer Science, pages 63–76, Warsaw, Poland. Academic Publishing House Exit.

Magdalena Derwojedowa, Maciej Piasecki, Stanisław Szpakowicz, Magdalena Zawisławska, and Bartosz Broda. 2008. Words, concepts and relations in the construction of Polish WordNet. In Attila Tanacs, Dora Csendes, Veronica Vincze, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the Global WordNet Conference*, pages 162–177, Seged, Hungary.

Christiane Fellbaum, editor. 1998. *WordNet — An Electronic Lexical Database*. MIT Press, Cambridge, MA.

W. Nelson Francis and Henry Kucera. 1964, revised and amplified 1979. Brown corpus manual. Internet.

Elżbieta Hajnicz. 2013a. Actualising lexico-semantic annotation of *Składnica* treebank to modified versions of source resources. in preparation.

Elżbieta Hajnicz. 2013b. Mapping named entities from NKJP corpus to *Składnica* treebank and polish wordnet. In Mieczysław A. Kłopotek, Jacek Koronacki, Małgorzata Marciniak, Agnieszka Mykowiecka, and Sławomir T. Wierzchoń, editors, *Proceedings of the 20th International Conference on Language Processing and Intelligent Information Systems*, volume 7912 of *LNCS*, pages 92–105, Warsaw, Poland. Springer-Verlag.

Elżbieta Hajnicz. 2013c. Procedure and results of the lexico-semantic annotation of *Składnica* treebank. in preparation.

Łukasz Kobyliński. 2011. Mining class association rules for word sense disambiguation. In Pascal Bouvry, Mieczysław A. Kłopotek, Franck Leprevost, Małgorzata Marciniak, Agnieszka Mykowiecka, and Henryk Rybiński, editors, *Proceedings of the International Joint Conference on Security and Intelligent Information Systems*, volume 7053 of *LNCS*, pages 307–317, Warsaw, Poland. Springer-Verlag.

2000. *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2000)*, Athens, Greece.

Diana McCarthy. 2001. *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences*. PhD thesis, University of Sussex.

George A. Miller and Christiane Fellbaum. 2007. WordNet then and now. *Language Resources and Evaluation*, 41:209–214.

George A. Miller, 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.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross Bunker. 1993. A semantic concordance. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 303–308, Plainsboro, NJ.

Simonetta Montemagni, Francesco Barsotti, Marco Battista, Nicoletta Calzolari, Ornella Corazzari, Alessandro Lenci, Vito Pirrelli, Antonio Zampolli, Francesca Fanciulli, Maria Massetani, Remo Raffaelli, Roberto Basili, Maria Teresa Pazienza, Dario Saracino, Fabio Zanzotto, Nadia Mana, Fabio Pianesi, and Rodolfo Delmonte. 2003a. The syntactic-semantic treebank of Italian. an overview. *Linguistica Computazionale*, XVI–XVI:461–492.

Simonetta Montemagni, Francesco Barsotti, Marco Battista, Nicoletta Calzolari, Ornella Corazzari, Alessandro Lenci, Antonio Zampolli, Francesca Fanciulli, Maria Massetani, Remo Raffaelli, Roberto Basili, Maria Teresa Pazienza, Dario Saracino, Fabio Zanzotto, Nadia Mana, Fabio Pianesi, and Rodolfo Delmonte. 2003b. Building the Italian syntactic-semantic treebank. In Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, Language and Speech, pages 189–210. Kluwer Academic Publishers, Dordrecht, Holland.

Martha Palmer, Hoa Trang Dang, and Joseph Rosenzweig. 2000. Semantic tagging the Penn treebank. In LREC (LRE, 2000), pages 699–704.

Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. 2009. *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław, Poland.

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, Barbara Lewandowska-Tomaszczyk, Marek Łaziński, and Piotr Pęzik. 2011. National Corpus of Polish. In Vetulani (Vetulani, 2011), pages 259–263.

Adriana Roventini, Antonietta Alonge, Nicoletta Calzolari, Bernardo Magnini, and Francesca Bertagna. 2000. ItalWordNet: a large semantic database for Italian. In LREC (LRE, 2000), pages 783–790.

Agata Savary, Jakub Waszczuk, and Adam Przepiórkowski. 2010. Towards the annotation of named entities in the National Corpus of Polish. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-2010)*, Valetta, Malta. ELRA.

Sabine Schulte im Walde. 2006. Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics*, 32(2):159–194.

Chaloemphon Sirkayon and Asanee Kawtrakul. 2007. Automatic lexico-semantic acquisition from syntactic parsed tree by using clustering and combining techniques. In *Proceedings of the International Workshop on Intelligent Systems and Smart Home (WISH 2007)*, volume 4743 of *LNCS*, pages 203–213. Springer-Verlag.

Marek Świdziński. 1992. *Gramatyka formalna języka polskiego*. Rozprawy Uniwersytetu Warszawskiego. Wydawnictwa Uniwersytetu Warszawskiego, Warsaw, Poland.

Marek Świdziński and Marcin Woliński. 2009. A new formal definition of Polish nominal phrases. In Małgorzata Marciniak and Agnieszka Mykowiecka, editors, *Aspects of Natural Language Processing*, volume 5070 of *LNCS*, pages 143–162. Springer-Verlag.

Marek Świdziński and Marcin Woliński. 2010. Towards a bank of constituent parse trees for Polish. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Proceedings of the International Conference on Text, Speech and Dialogue TSD 2010*, volume 6231 of *LNAI*, pages 197–204, Brno, Czech Republic. Springer-Verlag.

Zygmunt Vetulani, editor. 2011. *Proceedings of the 5th Language & Technology Conference*, Poznań, Poland.

Marcin Woliński, Katarzyna Głowińska, and Marek Świdziński. 2011. A preliminary version of Składnica — a treebank of Polish. In Vetulani (Vetulani, 2011), pages 299–303.

Marcin Woliński. 2005. An efficient implementation of a large grammar of Polish. In Zygmunt Vetulani, editor, *Proceedings of the 2nd Language & Technology Conference*, pages 343—-347, Poznań, Poland.