The Procedure of Lexico-Semantic Annotation of *Składnica* Treebank

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, the procedure of lexico-semantic annotation of *Składnica* Treebank using Polish WordNet is presented. Other semantically annotated corpora, in particular treebanks, are outlined first. Resources involved in annotation as well as a tool called *Semantikon* used for it are described. The main part of the paper is the analysis of the applied procedure. It consists of the basic and correction phases. During basic phase all nouns, verbs and adjectives are annotated with wordnet senses. The annotation is performed independently by two linguists. Multi-word units obtain special tags, synonyms and hypernyms are used for senses absent in Polish WordNet. Additionally, each sentence receives its general assessment. During the correction phase, conflicts are resolved by the linguist supervising the process. Finally, some statistics of the results of annotation are given, including inter-annotator agreement. The final resource is represented in XML files preserving the structure of *Składnica*.

Keywords: treebanks, wordnets, semantic annotation, Polish

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 as they have a number of 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). Semantically annotated treebanks are the important part of semantically annotated corpora.

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 making up sentences are annotated using Polish WordNet (PLWN, cf. section 3.2.) lexical units.

While elaborating the procedure, we were focused on applying the annotation for automatic support of semantic valence dictionary (in particular, establishing selectional preferences) and for training probabilistic semantic parsing. Using the treebank for WSD had a lower priority.

The annotation is performed using a dedicated tool called *Semantikon*. Each sentence is annotated by two linguists, and conflicts are resolved by a linguist supervising the process.

The procedure of lexico-semantic annotation of *Składnica* was preceded by semi-automatic tagging of named entities with corresponding PLWN-base semantic types (Hajnicz, 2013). Unlike with common words, this information was linked to nonterminal nodes, since named entities are very often multi-word units.

Section 2. presents related work on semantic annotation of text corpora. Section 3. contains the description of resources used. The applied procedure of the annotation of tokens is discussed in section 4., whereas section 5. contains some statistics of the results of annotation.

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. They are usually substantially smaller than other types of corpora.

The best known 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. It was annotated using a dedicated interface called ConText (Leacock, 1993). The corpus was preprocessed in order to find proper names and collocations (the ones present in PWN). The collocations were combined into single units by concatenating them with underscores (e.g., *took_place*). ConText analyses a corpus word by word (only open-class words). Annotators choose an appropriate sense from a list. They also have a possibility of adding comments when no available sense is considered appropriate.

A 1.7 mln subcorpus of the British National Corpus was manually semantically annotated during the lexicographic project Hector (Atkins, 1991). All occurrences of 300 word types having between 300 and 1000 occurrences in this subcorpus were tagged, resulting in 220 000 tagged tokens.

As for Slavic languages, most words in the balanced subcorpus of Russian National Corpus (RNC) (Grishina and Rakhilina, 2005) were semantically annotated. The semantic annotation (Apresjan et al., 2006; Lashevskaja, 2006; Kustova et al., 2007) is based on a hierarchical taxonomic classification of a Russian lexicon *Lexicograph* (Filipenko et al., 1992, www.lexicograph.ru). The texts were semantically tagged with a program named *Semmarkup* (elaborated by A. Polyakov).

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).

Unlike other corpora, semantic annotation of treebanks usually is not limited to lexico-semantic annotation. For example, PropBank (Kingsbury and Palmer, 2002; Kingsbury et al., 2002, `http//www.cis.wpenn.edu/ace`) constitutes the extension of the WSJ part of the Penn Treebank (Marcus, 1994; Marcus et al., 1993; Taylor et al., 2003) with predicate-argument structure. The French treebank TALANA (Abeillé et al., 1998; Abeillé et al., 2003) was annotated semantically using Dependency Minimal Recursion Semantics (Copestake, 2003) which contains predicate-argument relations, the restriction of generalised quantifiers, and combinations of predicates (Guillaume and Perrier, 2012).

Semantic annotation of Prague Dependency Treebank (Böhmová et al., 2003; Hajič, 2005) has the form of tectogrammatical dependency tree structures, where autosemantic words equipped with functors (deep roles) are semantically related.

Nevertheless, there exist some lexico-semantically annotated treebanks. First, a fragment of the Penn Treebank (Taylor et al., 2003) was lexico-semantically tagged with PWN senses (Palmer et al., 2000). Only the verbs and semantic heads of their arguments and adjuncts were annotated. In addition, proper nouns absent in PWN were tagged as either person, company, date or other name. Wherever possible, pronouns were tagged with the sense of their antecedents. The annotation was performed by two annotators, the second one making their own and the final decision. The inter-annotator agreement measured as percentage of consistent independent labelling was 89%. Special tags were used for lack of a proper sense in PWN and for uncertainty as to which sense is correct in a particular context.

The Portuguese Treebank *Floresta sintá(c)tica* (Alfonso et al., 2002) was annotated using a predefined hierarchy of semantic tags called *semantic prototypes* (Bick, 2006).

An interesting example is made by the Italian Syntactic-Semantic Treebank (Montemagni et al., 2003b; Montemagni et al., 2003a), which includes a functional level of annotation (which is a borderline between syntax and semantics) and a lexico-semantic level of annotation. Its 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 metonymic usage of words or expressions, diminutive and augmentative derivatives, and idioms. Moreover, named entities are tagged with their (rather coarse) semantic types. The lexico-semantic level of annotation is separated from the syntactic one; both refer to the morphosyntactic level.

3. Data Resources

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

```
<node nid="48" from="7" to="8" chosen="true">
    <nonterminal>
        <category>formarzecz</category>
    </nonterminal>
    <children rule="n_rz1" chosen="true">
        <child nid="49" from="7" to="8"/>
    </children>
</node>
<node nid="49" from="7" to="8" chosen="true">
    <terminal token_id="morph_6.75-seg">
        <orth>pokoleń</orth>
        <base>pokolenie</base>
        <f type="tag">subst:pl:gen:n</f>
    </terminal>

<plwn_interpretation .../>

</node>
```

Figure 1: Fragment of the representation of a sentence in *Składnica*

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.

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 was accepted as correct.

Parse trees are encoded in XML, each parse being stored in a separate file. Each tree node, terminal or nonterminal, is represented by means of an XML `node` element, having two attributes `from` and `to` determining the boundaries of the corresponding phrase. Terminals additionally contain a `token_id` attribute linking them with corresponding NKJP tokens.

A fragment of the representation of sentence (1) in *Składnica* is shown in Fig. 1.

(1) *Taki był u nas zwyczaj od pokoleń.*
such was among us habit for generations
There was such a habit among us for generations.

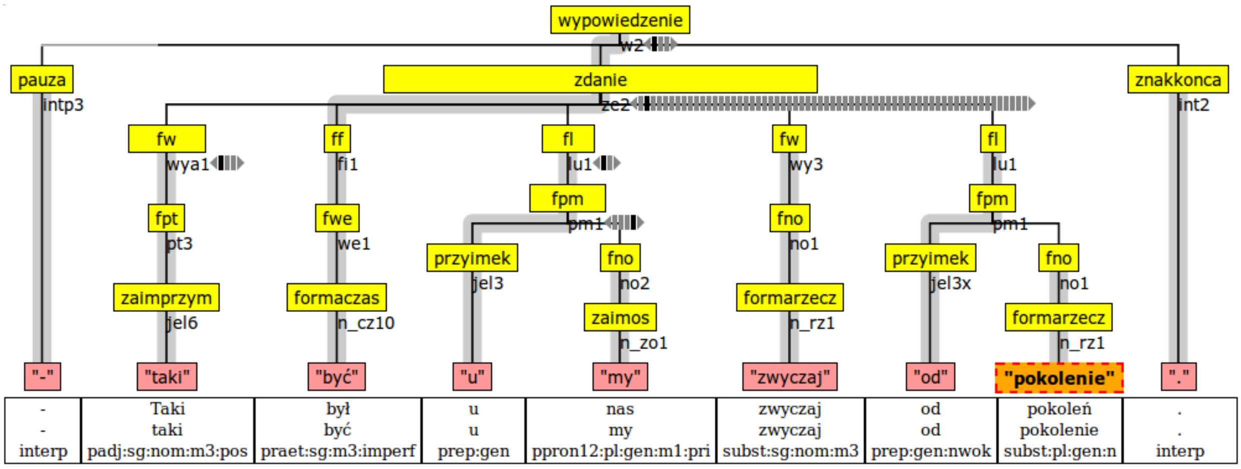
The parse tree of sentence (1) in *Składnica* is shown in Fig. 2. Thick gray shadows emphasising some branches in the tree show heads of phrases.

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 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

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



-	Taki	był	u	nas	zwyczaj	od	pokoleń	.
-	taki	być	u	my	zwyczaj	od	pokolenie	.
interp	padj:sg:nom:m3:pos	praet:sg:m3:imperf	prep:gen	ppron12:pl:gen:m1:pri	subst:sg:nom:m3	prep:gen:nwok	subst:pl:gen:n	interp

Figure 2: Exemplary parse tree from *Składnica*

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, an 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.

Version 2.0 of PLWN is used for the semantic annotation of tokens. On the other hand, named entity annotation was performed by means of PLWN 1.6.

4. Main Rules 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. 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 with parse trees are annotated. The reason is that corresponding LUs are assigned to terminal nodes representing tokens being annotated. This feature can limit applicability of the resulting resource in WSD.

4.2. XML Encoding

Semantic annotation is introduced into the XML structure of a parse tree as a new type of a child element of the element `node`: a terminal node for common words and a nonterminal node for named entities. All corresponding LUs are included, the correct one having attribute `chosen="true"` (see Fig. 3).

Additionally, the root element is augmented with three attributes, `name-plwn_version`, `sense-plwn_version` and `final-plwn_version` pointing out which version of PLWN was used for a particular phase of semantic annotation.

4.3. Organisation of Annotation

Annotation is performed by means of a dedicated tool called *Semantikon*. A tree of directories of *Składnica* XML files grouped into paragraphs together with XML version of PLWNconstitutes its input. During the process of annotation they are stored as a MySQL database, together with information concerning users, etc. The results of semantic annotation are written into augmented XML files organised in the same way.

```
<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

The program is written in `Java JDK` 1.6 as application server technology and run in `Tomcat` v.6.2 environment using `Hibernate` technology. The interface is encoded with `Java Server Pages JSP` 2.0 and `Java Server Faces JSF` technology.

We have identified the following roles of users of the system:

- **annotator** — selects corresponding LUs for each token,
- **supervisor** — resolves conflicts between annotators,
- **administrator** — maintains the system database.

Users access the system via an Internet browser.

We decided to implement a dedicated tool, as we expected the tool to provide some specific functionalities. Namely,

- the whole paragraph visible on a screen,
- possibility of special tagging and adaptation of the list of PLWN LUs to be chosen for the particular way of tagging,
- assessment of a whole sentence,
- user management (including salaries).

4.3.1. Basic Annotation Phase

The linear (or "textual") method was chosen for annotation. In our opinion the transversal (or "lexical") method consisting in tagging all occurrences of a lexeme together is better suited to *lexical sample* tasks, hence it is too time consuming for *all words* tasks.[2]

Users are provided with data in portions corresponding to single paragraphs, represented as lists of sentences (see the top table of Fig. 4). The column `Status` presents the phase of annotation of the sentence: waiting for annotation (green ✓), under annotation (blue ●), accepted (violet ●), rejected (violet ✗). Sentences marked with a green ✓ in the `OK` column have a proper parse in *Składnica* and are intended for annotation. Adjectives, nouns, and verbs in a sentence chosen for annotation are marked in the same way (see the bottom left table of Fig. 4).

For each annotated token, the list of PLWN LUs with a corresponding lemma is presented (the right bottom table of Fig. 4). The user can access more information about each LU available by PLWN browser. After clicking the `Akceptuj` ('accept') button the identifier of the current LU appears in the `Interp` column.

If more than one LU seems adequate for a token in the context the annotator is asked to choose the LU with the least sense number (supposed to be most frequent) among the adequate ones. This raises the level of inter-annotation agreement, but we loose the information that the token was hard for annotation.

The user is provided with several auxiliary buttons for the case when no PLWN LU is adequate for a token. First, in Polish the reflexive marker *się* can be the part of the verb lemma (e.g., *uśmiechać się*, 'to smile'). Thus, clicking the `zawiera 'się'` ('contains reflexive marker') button causes the display of the list of LUs with the verb lemma plus *się* (active only for verb tokens in presence of the reflexive marker in the sentence).

The user can also choose the corresponding multi-word expression present in PLWN. The list of multi-word units containing a particular token lemma becomes available after clicking the corresponding button. What is important, we always annotate single tokens. In the case of a multi-word expression it is its semantic head (other tokens are marked with an auxiliary tag as elements of a multi-word expression).

The other option is to choose a synonym or a hypernym of a token. In such a case a user is asked to type its lemma. The special tag `Brak` ('lack') is used for the case when a synonym or even a hypernym of a lacking LU is hard to establish (in contrast to the Penn treebank annotation).

```
<node nid="5" from="0" to="1" chosen="true">
    <terminal token_id="morph_1.42-seg" disamb="true">
        <orth>Denat</orth>
        <base>denat</base>
        <f type="tag">subst:sg:nom:m1</f>
    </terminal>
    <plwn_interpretation sem_id="sem_1">
        <plwn_units case_agreement="true"
                    polysemy="true">
            <unit luid="sem_1-sv1" chosen="true">
                <lubase>denat</lubase>
                <lusense>1</lusense>
                <luident>48485</luident>
                <synset>31397</synset>
            </unit>
            .................................
        </plwn_units>
    </plwn_interpretation>
</node>
.............................................
<node nid="19" from="4" to="5" chosen="true">
    <terminal token_id="morph_1.46-seg" disamb="true">
        <orth>sobie</orth>
        <base>siebie</base>
        <f type="tag">siebie:loc</f>
    </terminal>
    <plwn_interpretation sem_id="sem_3" type="anaphora">
        <anaphora ref_sent="self" ref_node="5"/>
    </plwn_interpretation>
</node>
```

Figure 5: XML representation of an anaphoric usage of a pronoun

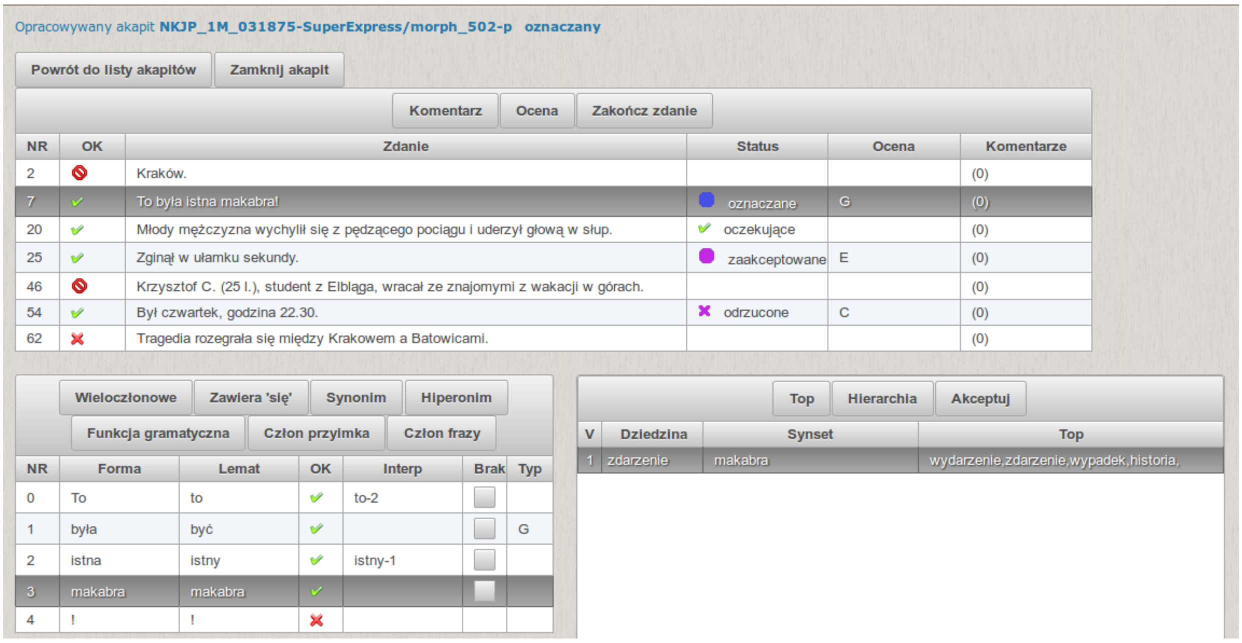
```
<node nid="50" from="0" to="1" chosen="true">
   <terminal token_id="morph_81.1-seg" disamb="true">
      <orth>On</orth>
      <base>on</base>
      <f type="tag">ppron3:sg:nom:m1:ter:akc:npraep</f>
   </terminal>
   <plwn_interpretation sem_id="sem_3" type="anaphora">
      <anaphora ref_sent="context" ref_node="none">
         <unit luid="ana_3-sv1">
             <lubase>mężczyzna</lubase>
             <lusense>1</lusense>
             <luident>3419</luident>
              <synset>6709</synset>
         </unit>
      </anaphora>
   </plwn_interpretation>
</node>
```

Figure 6: XML representation of a deictic or undetermined usage of a pronoun

Other buttons are used to indicate tokens which for some reasons need not be annotated. In particular, 3rd person (sentence (2)), self (sentence (3)), and WH- (sentence (4)) pronouns are tagged as anaphoric[3] and linked with their antecedent (by pointing the corresponding sentence and node identifiers, see Fig. 5 for sentence (3)). This preserves the consistency of annotation regardless of changes in tagging. Since the context is limited to a single paragraph, very often we are not able to distinguish a deictic use of a pronoun from anaphoric one with the antecedent in a longer context. Anyway, in such cases, the word `context` is used instead of an antecedent, and the sense is introduced by means of a corresponding LU introduced by the annotator (in similar way to synonyms and hypernyms). For example, the sentences in (5) start the paragraph. However, the context shows us that sentences refer to a man (*mężczyzna*), cf. Fig. 6.

[2] Nevertheless, it is sometimes used for it, e.g., (Navarro et al., 2005).

[3] 1st and 2nd person pronouns are automatically tagged with LU osoba-1 ('person').


Opracowywany akapit **NKJP_1M_031875-SuperExpress/morph_502-p** **oznaczany**

NR	OK	Zdanie	Status	Ocena	Komentarze
2	🚫	Kraków.			(0)
7	✓	To była istna makabra!	● oznaczane	G	(0)
20	✓	Młody mężczyzna wychylił się z pędzącego pociągu i uderzył głową w słup.	✓ oczekujące		(0)
25	✓	Zginął w ułamku sekundy.	● zaakceptowane	E	(0)
46	🚫	Krzysztof C. (25 l.), student z Elbląga, wracał ze znajomymi z wakacji w górach.			(0)
54	✓	Był czwartek, godzina 22.30.	✗ odrzucone	C	(0)
62	✗	Tragedia rozegrała się między Krakowem a Batowicami.			(0)

NR	Forma	Lemat	OK	Interp	Brak	Typ
0	To	to	✓	to-2	□	
1	była	być	✓		□	G
2	istna	istny	✓	istny-1	□	
3	makabra	makabra	✓		□	
4	!	!	✗			

V	Dziedzina	Synset	Top
1	zdarzenie	makabra	wydarzenie,zdarzenie,wypadek,historia,

Figure 4: Annotation interface of a particular sentence in a paragraph

(2) *Jenny$_1$ przespała granicę i nikt **jej**$_1$ nie budził.*
Jenny overslept border and nobody her not waked
Jenny overslept the border and nobody waked her up.

(3) *Denat$_1$ nie miał przy **sobie**$_1$ dokumentów.*
deceased not had with self documents
The deceased had not documents with him.

(4) *Obok leżał rower$_1$, **którym**$_1$ jechała kobieta.*
next lay bicycle that rode woman
A bicycle the woman rode was lying next to her.

(5) a. ***On** również myśli o przyszłości.*
He also thinks about future
He thinks also about the future.

b. *Dlatego skończył polonistykę*
That's why graduated Polish studies
i dziennikarstwo.
and journalism.
That's why he graduated from Polish studies and journalism.

In the case of tokens being elements of multi-words 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, although they do not make the majority.
In order to finish annotation of a sentence, the user should assess it. The following assessments are available:

1. fully annotated sentence (F),
2. lack of corresponding lemma (L),
3. lack of corresponding LU (J),
4. occurrence of anaphora (A),
5. occurrence of ellipsis (E),
6. occurrence of metaphor (M),
7. occurrence of metonymy (H),
8. incorrect lemmatisation of a token (I),
9. incorrect sentence (C).

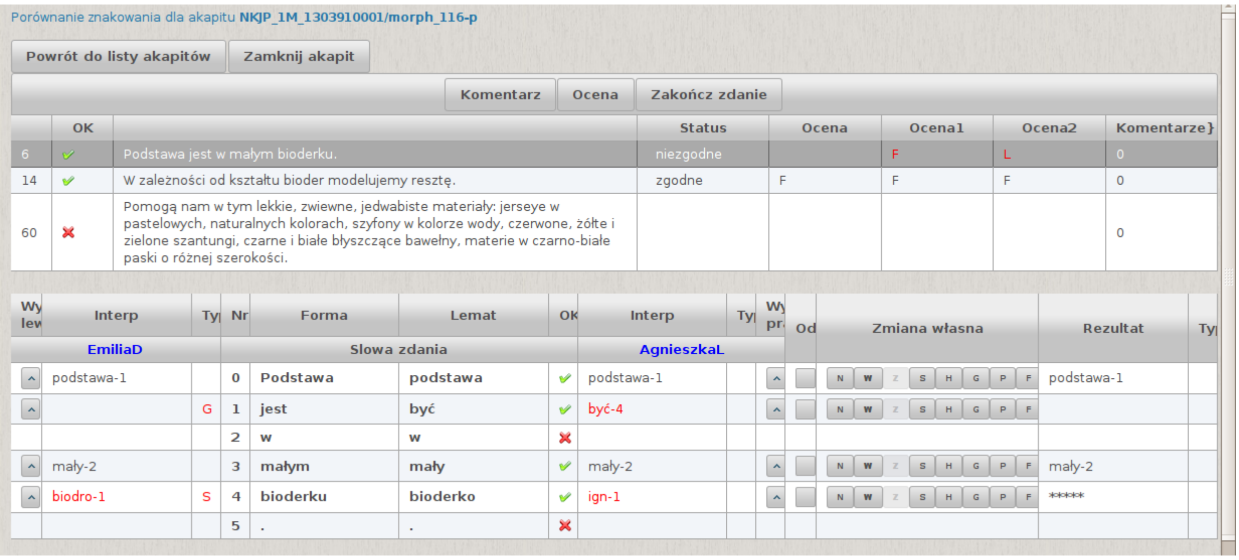
The first assessment requires that the annotation of all autosemantic tokens in the sentence is correct and final, the last one means that the sentence was considered incorrect and was not annotated at all. Other marks concern particular problems and phenomena occurring in the sentence, hence the annotator can choose as many of them as they find appropriate for a particular sentence (a list of assessments). 2nd and 3rd marks indicate lacks in PLWN, 8th mark indicates an incorrect lemma (e.g., in *odwracała ode mnie odwagę*, '[she] distracted courage [attention] from me', the word *odwaga* 'courage' is used instead of *uwaga* 'attention'). Only non-lexicalised usages of metaphor (6th mark) and metonymy (7th mark) are considered, e.g., *Do Polski kapitalizm wjechał czołgiem i kompletnie nas staranował.* ('Capitalism drove into Poland on a tank and completely rammed us').
Our goal to apply the treebank to support the valence dictionary creation was the reason of marking of ellipsis (5th mark). Therefore, we consider two types of ellipsis

- verb argument absence, including zero subjects (sentence (5) b),
- NP verb argument realised by AdjP or NumP (sentence (6)).

(6) *W tym roku prawdziwa zima$_1$*
At this year real winter
wyprzedziła kalendarzową$_{[\mathrm{el}]_1}$.
overtook calendar$_{adj}$
This year the real winter overtook the winter by the calendar.

4.3.2. Correction Phase

Following the common practice, each sentence in *Semantikon* is processed by the two annotators independently. If they assign the same LU to every token being annotated and



Porównanie znakowania dla akapitu **NKJP_1M_1303910001/morph_116-p**

	OK		Status	Ocena	Ocena1	Ocena2	Komentarze}
6	✓	Podstawa jest w małym bioderku.	niezgodne		F	L	0
14	✓	W zależności od kształtu bioder modelujemy resztę.	zgodne	F	F	F	0
60	✗	Pomogą nam w tym lekkie, zwiewne, jedwabiste materiały: jerseye w pastelowych, naturalnych kolorach, szyfony w kolorze wody, czerwone, żółte i zielone szantungi, czarne i białe błyszczące bawełny, materie w czarno-białe paski o różnej szerokości.					0

Wy lew	Interp	Ty	Nr	Forma	Lemat	OK	Interp	Ty	Wy pr	Od	Zmiana własna	Rezultat	Ty
EmiliaD			Słowa zdania				AgnieszkaL						
^	podstawa-1		0	Podstawa	podstawa	✓	podstawa-1		^	□	N W Z S H G P F	podstawa-1	
^		G	1	jest	być	✓	być-4		^	□	N W Z S H G P F		
			2	w	w	✗							
^	mały-2		3	małym	mały	✓	mały-2		^	□	N W Z S H G P F	mały-2	
^	biodro-1	S	4	bioderku	bioderko	✓	ign-1		^	□	N W Z S H G P F	*****	
			5	.	.	✗							

Figure 7: Supervisor interface of a sentence with a conflict of annotation

give the same set of assessments to the sentence as a whole, it is considered valid (although the supervisor can inspect and change it). If any difference is detected by the system, each of the two annotators is asked to check their annotation. They can see which tokens were annotated differently by the other annotator but they do not know the other answer. After this procedure, if the annotations of the sentence still differ, the sentence is marked as a conflict and passed to the supervisor who will accept one of the annotations or give a new one. The supervisor sees both annotations, with the differences marked in red (cf. Fig. 7). The choices of LUs made by each annotator are stored in XML files (attribute `chosen="single"`).

5. Results of Annotation

The procedure of semantic annotation of *Składnica* is not finished yet. The version 0.5 of *Składnica* to be annotated consists of 8283 manually validated trees containing 49264 nouns, verbs, and adjectives for annotation. 6970 sentences (84%) including 39005 words have undergone basic annotation phase, whereas 6135 (74%) sentences including 34270 words have undergone the whole correction phase.

While annotating sentences and correcting possible small errors, the annotators made the same decision for 28984 (84%) words, which is worse than 89% reported by (Palmer et al., 2000) and better than 76% reported by (Navarro et al., 2005) for Spanish corpus annotated with Spanish WordNet sense inventory. In particular, 76 (16%) synonyms and 128 (24%) hypernyms were chosen identical. Moreover, 14238 (72%) of 19708 polysemic words annotated in standard way (including reflexive verbs) were annotated consistently. However, the supervisor decided to intervene in 5835 (17%) cases, which means that she changed 549 consistent annotations. The number of sentences that did require supervisor's intervention (including general assessments) is 4079 (66%).

For 34237 finally annotated words, 28288 (82%) have standard annotation, 475 were annotated with a synonym, 527 with a hypernym, and 352 with a multi-word expression. 839 verbs were annotated as reflexive. Moreover, 856 tokens were considered as such that do not need to be annotated for some reasons, cf. (Hajnicz, 2014), whereas for 2910 tokens (8.5%) no adequate annotation was found.

For 6130 finally annotated sentences, 5 were considered incorrect, whereas 96 include improperly lemmatised tokens. Anaphora and ellipsis were detected in 1232 and 1031 sentences, respectively.

Some lacks of sense concern tokens lemmatised improperly in *Składnica*, neologisms etc. Nevertheless, most of indicate lacks in PLWN: there was no corresponding lexical unit for 638 tokens, whereas no lemma of 263 tokens was found in PLWN.

Finally, 99 sentences constitute a metaphoric usage and 24 constitutes a metonymic one.

6. Conclusions and Future Work

In this paper, the procedure of the ongoing work concerning lexico-semantic annotation of a treebank was presented. The possibility of annotating tokens by means of synonyms and hypernyms is the main novelty of our approach. This makes the method more complicated, as the annotators are to some extent free to choose the LU adequate for a particular token. Therefore, 84% of consistent annotations seems to be a very good result, the more so as choosing the same synonym or hypernym is very unlikely. It was obtained due to the decision to choose the least sense number of a word in the case of ambiguity.

In the future we plan to annotate ellipses (representing their morphosyntactic characteristics and position in a parse tree) and link them with their anaphoric antecedents, as they inherit their lexico-semantic interpretation. This information is required for extracting semantic valence from parse trees.

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.

7. References

Abeillé, A., editor. (2003). *Treebanks: Building and Using Parsed Corpora*. Language and Speech. Kluwer Academic Publishers, Dordrecht, Holland.

Abeillé, A., Clément, L., and Reyes, R. (1998). TALANA annotated corpus for French: the first results. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC-1998)*, pages 992–999, Grenada, Spain.

Abeillé, A., Clement, L., and ois Toussenel, F. (2003). Building a treebank for French. In (Abeillé, 2003), pages 165–187.

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

Alfonso, S., Bick, E., Haber, R., and Santos, D. (2002). Floresta sintá(c)tica: a treebank of portuguese. In (LREC, 2002), pages 1698–1703.

Apresjan, J., Boguslavsky, I., Iomdin, L., Iomdin, B., Sannikov, A., and Sizov, V. (2006). A syntactically and semantically tagged corpus of Russian: State of the art and prospects. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*, pages 1378–1381, Genoa, Italy.

Atkins, S. (1991). Tools for computer-aided corpus lexicography: The Hector project. *Acta Linguistica Hungarica*, 41:5–72.

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

Böhmová, A., Hajičová, E., Hajič, J., and Hladká, B. (2003). The Prague dependency treebank: A three-level annotation scenario. In (Abeillé, 2003), pages 103–127.

Broda, B., Piasecki, M., and Maziarz, M. (2009). Evaluating LexCSD—a weakly-supervised method on improved semantically annotated corpus in a large scale experiment. In Kłopotek, M. A., Marciniak, M., Mykowiecka, A., Penczek, W., and Wierzchoń, S. T., editors, *Intelligent Information Systems*, Challenging Problems in Science: Computer Science, pages 63–76, Warsaw, Poland. Academic Publishing House Exit.

Copestake, A. (2003). Slacker semantics: Why superficiality, dependency and avoidance of commitment can be the right way to go (invited talk). In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2003)*, pages 1–9, Budapest, Hungary.

Derwojedowa, M., Piasecki, M., Szpakowicz, S., Zawisławska, M., and Broda, B. (2008). Words, concepts and relations in the construction of Polish WordNet. In Tanacs, A., Csendes, D., Vincze, V., Fellbaum, C., and Vossen, P., editors, *Proceedings of the Global WordNet Conference*, pages 162–177, Seged, Hungary.

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

Filipenko, M., Paducheva, E., and Rakhilina, E. (1992). Semantic dictionary viewed as a lexical database. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-1992)*, pages 1295–1299, Nantes, France.

Francis, W. N. and Kucera, H. (1964, revised and amplified 1979). Brown corpus manual. Internet.

Grishina, E. and Rakhilina, E. (2005). Russian National Corpus (RNC): an overview and perspectives. In *Proceedings of the AATSEEL 2005*.

Guillaume, B. and Perrier, G. (2012). Semantic annotation of the french treebank with modular graph rewriting. In *Proceedings of META-RESEARCH Workshop on Advanced Treebanking, LREC 2012*, pages 14–21, Istanbul, Turkey.

Hajič, J. (2005). Complex corpus annotation: The Prague dependency treebank. In Šimková, M., editor, *Insight into Slovak and Czech Corpus Linguistics*, pages 54–73. Veda, Bratislava, Slovakia.

Hajnicz, E. (2013). Mapping named entities from NKJP corpus to *Składnica* treebank and Polish WordNet. In Kłopotek, M. A., Koronacki, J., Marciniak, M., Mykowiecka, A., and Wierzchoń, S. T., 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.

Hajnicz, E. (2014). Lexico-semantic annotation of *Składnica* treebank by means of PLWN lexical units. In Orav, H., Fellbaum, C., and Vossen, P., editors, *Proceedings of the 7th International WordNet Conference (GWC 2014)*, pages 23–31, Tartu, Estonia. University of Tartu.

Kingsbury, P. and Palmer, M. (2002). From TreeBank to PropBank. In (LREC, 2002), pages 1989–1993.

Kingsbury, P., Palmer, M., and Marcus, M. P. (2002). Adding semantic annotation to the Penn TreeBank. In *Proceedings of the Human Language Technology Conference*, pages 252–256, San Diego, CA.

Kobyliński, Ł. (2011). Mining class association rules for word sense disambiguation. In Bouvry, P., Kłopotek, M. A., Leprevost, F., Marciniak, M., Mykowiecka, A., and Rybiński, H., editors, *Proceedings of the International Joint Conference on Security and Intelligent Information Systems*, volume 7053 of *LNCS*, pages 307–317, Warsaw, Poland. Springer-Verlag.

Kustova, G., Lashevskaja, O., Rakhilina, E., and Paducheva, E. (2007). On taxonomy in cognitive semantics and corpus linguistics: Parts of body. In *Proceedings of the 10th International Cognitive Conference*, Cracow, Poland.

Lashevskaja, O. (2006). Corpus-aided construction grammar: Semantic tools in the Russian National Corpus. In *Proceedings of the 2th International Meeting of the German Cognitive Linguistic Association*, Munich, Germany.

Leacock, C. (1993). Context: A toot for semantic tagging of text: User's guide. Technical Report 54, Cognitive Science Laboratory, Princeton University, Princeton, NJ.

LREC (2000). *Proceedings of the 2nd International Con-*

ference on Language Resources and Evaluation (LREC-2000), Athens, Greece.

LREC (2002). *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, Las Palmas, Spain.

Marcus, M. P. (1994). The Penn TreeBank: A revised corpus design for extracting predicate-argument structure. In *Proceedings of the ARPA Human Language Technology Workshop*, Princeton, NJ. Morgan Kaufmann.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

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

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to WordNet: An online lexical database. *International Journal of Lexicography*, 3(4):235–244.

Miller, G. A. and Fellbaum, C. (2007). WordNet then and now. *Language Resources and Evaluation*, 41:209–214.

Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. (1993). A semantic concordance. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 303–308, Plainsboro, NJ.

Montemagni, S., Barsotti, F., Battista, M., Calzolari, N., Corazzari, O., Lenci, A., Pirrelli, V., Zampolli, A., Fanciulli, F., Massetani, M., Raffaelli, R., Basili, R., Pazienza, M. T., Saracino, D., Zanzotto, F., Mana, N., Pianesi, F., and Delmonte, R. (2003a). The syntactic-semantic treebank of Italian. An overview. *Linguistica Computazionale*, XVI–XVI:461–492.

Montemagni, S., Barsotti, F., Battista, M., Calzolari, N., Corazzari, O., Lenci, A., Zampolli, A., Fanciulli, F., Massetani, M., Raffaelli, R., Basili, R., Pazienza, M. T., Saracino, D., Zanzotto, F., Mana, N., Pianesi, F., and Delmonte, R. (2003b). Building the Italian syntactic-semantic treebank. In (Abeillé, 2003), pages 189–210.

Navarro, B., Marcos, R., and Abad, P. (2005). Semantic annotation and inter-annotator agreement in cast3lb corpus. In Civit, M., Kübler, S., and Martí, M. A., editors, *Proceedings of the 4th Workshop on Treebanks and Linguistic Theories*, pages 125–135, Barcelona, Spain.

Palmer, M., Dang, H. T., and Rosenzweig, J. (2000). Semantic tagging for the Penn Treebank. In (LREC, 2000), pages 699–704.

Piasecki, M., Szpakowicz, S., and Broda, B. (2009). *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław, Poland.

Przepiórkowski, A., Bańko, M., Górski, R. L., Lewandowska-Tomaszczyk, B., Łaziński, M., and Pęzik, P. (2011). National Corpus of Polish. In (Vetulani, 2011), pages 259–263.

Roventini, A., Alonge, A., Calzolari, N., Magnini, B., and Bertagna, F. (2000). ItalWordNet: a large semantic database for Italian. In (LREC, 2000), pages 783–790.

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

Sirkayon, C. and Kawtrakul, A. (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.

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

Świdziński, M. (1992). *Gramatyka formalna języka polskiego*. Rozprawy Uniwersytetu Warszawskiego. Wydawnictwa Uniwersytetu Warszawskiego, Warsaw, Poland.

Świdziński, M. and Woliński, M. (2010). Towards a bank of constituent parse trees for Polish. In Sojka, P., Horák, A., Kopeček, I., and Pala, K., 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.

Taylor, A., Marcus, M. P., and Santorini, B. (2003). The Penn Treebank: An overview. In (Abeillé, 2003), pages 5–22.

Vetulani, Z., editor. (2011). *Proceedings of the 5th Language & Technology Conference*, Poznań, Poland. Fundacja Uniwersytetu im. A. Mickiewicza.

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

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