Elżbieta Hajnicz, Institute of Computer Science, Polish Academy of Sciences, Warsaw

Towards Extending Syntactic Valence Dictionary for Polish with Semantic Categories

1. Introduction

The main goal of our work is to extend a syntactic valence dictionary of Polish verbs by adding semantic information, represented by means of wordnet *semantic categories* of nouns. Syntactic valence dictionary is a collection of *predicates* (here: verbs) provided with a set of *verb frames*. Each verb frame is composed of its *syntactic slots* representing phrases occurring in the corresponding position in a sentence. Thus, our goal is to provide syntactic slots (here: NPs/PPs) with a list of appropriate semantic categories of corresponding nouns.

In this paper, we want to emphasize various problems and obstacles that impede an automatic creation of such a dictionary for Polish.

A number of resources and tools necessary for Natural Language Processing (NLP) are already available for Polish: e.g., *morphological analysers* (or *dictionaries*) (Hajnicz and Kupść 2001; Rabiega-Wiśniewska 2004; Woliński 2006), *deep* (Obrębski 2002; Przepiórkowski et al. 2002; Woliński 2004) and *shallow* (Przepiórkowski 2007a, b) *parsers*. Deep parsers often use syntactic valence dictionaries (Mędak 2005; Świdziński 1994; Przepiórkowski 2006). However, recognising a syntactic structure of texts turns out to be insufficient to obtain satisfactory results in solving NLP tasks such as *machine translation*, *information extraction*, *question answering* — here semantic information is indispensable.

In practical applications focused on specific domains (e.g., medicine, finance, sport) such information is often gathered using *ontologies*. On the other hand, more universal lexical semantic resources, such as wordnets (Fellbaum 1998; Vossen 1998) and FrameNet (Baker et al. 1998; Boas 2002) are also created.

In order to automatically add semantic information to a syntactic valence dictionary, we need a treebank with all nouns semantically annotated with *categories* mentioned above, as both syntactic (i.e., argument structure) and semantic information is required.

We aim here at *Word Sense Disambiguation* (WSD). An extensive overview of the problem is presented in Agirre and Edmonds (2006). Methods used for WSD can be divided into *supervised,* which rely on a manually annotated subcorpus to train the algorithm (Abney 2004; Suárez and Palomar 2002), *unsupervised,* based on clustering words occurring in similar context rather than assigning senses from a given repertoire (Lin and Pantel 2002; Schütze 1998), and *knowledge-based*, applying electronic lexicons and lexical knowledge bases (such as wordnets) (Banerjee and Pedersen 2002; McCarthy and Carroll 2003). Most of these techniques are focused on fairly fine-grained word senses, hence they are applied to a small set of words or they need a big corpus to operate on.

As for Slavic languages, the only WSD method we are aware of was proposed for Czech (Král 2001). To the best of our knowledge, all other approaches consider a multilingual environment (Ion and Tufiş 2004). Actually, this is the main setup for WSD discussed in the literature. For Polish WSD is absolutely a new topic.

In order to create a semantic valence dictionary, we need sense annotations for words which are immediate arguments of a verb (heads of phrases). Therefore, we apply syntactic information (valence of the main verb in a clause) we have at our disposal to solve a WSD task. Unfortunately, such information is used to perform this task very rarely (Gaustad 2004). The frequently cited work in which syntactic information is used is Dorr and Jones (1996). However, they disambiguate only the semantic category of a verb (based on Levin (1993) classes) using a set of possible verb classes on one hand, and a syntactic frame of a sentence, on the other. Semantic categories of arguments are not considered.

In section 2, resources necessary for our approach are listed. In section 3, preparatory steps for data processing are presented. In section 4 we discuss morphosyntactic and syntactic

phenomena characteristic for Polish that may present problems for the method. In section 5 we show how the algorithm works for a small set of sentences containing a noun with two senses. Finally, we sum up the algorithm and present our future plans in section 6.

All examples presented throughout the paper are real sentences taken from the corpus used in the experiments.

2. Resources

First, a group of verbs for our experiments has been selected (Hajnicz 2007). It has been chosen manually in order to maximise the variability of syntactic frames (in particular, diathesis alternations) on one hand, and polysemy of verbs within a single syntacit frame, on the other. The frequency was an important criterion for this choice as well. The frequency of 99 selected verbs varies from 489 for *kupować* 'to buy' to 82 291 for *prosić* 'to ask for'.

Our main resource is a language corpus; we used the IPI PAN Corpus (Przepiórkowski 2004), a set of Polish written texts, segmented into paragraphs and sentences, annotated with morphosyntactic tags. The second edition contains 250 mln segments (roughly, words). From this corpus, we have selected a small subcorpus of 165,253 simple sentences containing the selected verbs. "Simple" means here containing just one verb. This was the only criterion to determine "simplicity" of sentences without information about their entire syntactic structure. All sentences considered in the subcorpus contain exclusively nouns present in Polish wordnet (see below). The subcorpus contains all sentences satisfying the above criteria.

Nr	symbol	name	Nr	symbol	name
001	bhp	Tops	014	cel	motive
002	czy	act	015	rz	object
003	zwz	animal	016	os	person
004	wytw	artifact	017	zj	phenomenon
005	cech	attribute	018	rsl	plant
006	czc	body	019	pos	possession
007	umy	cognition	020	prc	process
008	por	communication	021	il	quantity
009	zdarz	event	022	zw	relation
010	czuj	feeling	023	ksz	shape
011	jedz	food	024	st	state
012	grp	group	025	sbst	substance
013	msc	location	026	czas	time

Table 1: Predefined set of general semantic categories of nouns in Polish WordNet

In order to prepare an initial sense annotation for nouns (to be automatically disambiguated), we used the Polish WordNet (Derwojedowa et al. 2007; Zawisławska and Derwojedowa 2008), called *Słowosieć*. *Słowosieć* is a network of lexical-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 of one or more words by placing this unit in a network of links which represent such relations as synonymy, hypernymy, meronymy, etc. For the present work we do not use the whole structure of the net, but the set of 26 predefined categories (see table 1) which are at the top of the actual hierarchy. Using these categories, 7,815 nouns (most frequent in the balanced subcorpus of the IPI PAN Corpus) were classified by the Polish WordNet group. Each noun is assigned one to six categories.

3. Data Preparation

The selected subcorpus of 165,253 sentences mentioned above has been parsed with the metamorphic grammar *Świgra* (Woliński 2004), each parse reduced to its flat form identifying only the top-most phrases. The grammar takes sentences as they were identified in the corpus, but it ignores the disambiguation of morphosyntactic annotation as it was established by the tagger (i.e., it has taken into account all tags produced by the morphosyntactic analyser). Each phrase has a syntactic and a semantic head (cf. Przepiórkowski 2006). Usually, these heads are equal (i.e., they are heads of phrases determined by the grammar), but, for instance, the syntactic head of a preposition phrase is a preposition whereas its semantic head is a noun (i.e., the head of a complement noun phrase).

Next, a reduced parse forest for each sentence has been disambiguated by EM algorithm proposed for extraction of syntactic valence frames, the *EM selection algorithm* (Dębowski 2007; Dębowski and Woliński 2007). As a side-effect of this process we have obtained a syntactic valence dictionary. In particular, every verb token is associated with its corresponding syntactic frame.

Unfortunately, each of the applied tools has its own limitations concerning acceptable sentences, which results in reduction of the subcorpus. First, the present version of *Świgra* analyses only a subset of Polish syntactic constructions. Next, the EM selection algorithm discards sentences that have more than 40 reduced parses. Finally, we delete all sentences containing no NPs/PPs (such as *Ogromnie się cieszę.* '[I] am very glad.').[1] As a result, after taking into account these constraints, the number of sentences has dropped to 41,793. The statistics about sentences, their reduced parses and phrases they include are presented in table 2.

(*number of source sentences: 165 253*)	after syntactic analysis	after parse disambiguation	sentences with NPs/PPs only
Number of sentences	82 318	43 908	41 793
Number of reduced parses	2 368 531	57 328	53 065
Number of reduced parses per sentence	28.949	1.310	1.270
Number of phrases	9 394 185	120 786	114 320
Number of NPs and PPs	6 737 920	92 859	89 225
Average number of phrases per parse	3.966	2.107	2.154
Average number of NPs and PPs per parse	2.848	1.635	1.681

Table 2: Simple statistics about sentences and their parses

An important problem here is that the EM selection algorithm was developed for construction of a syntactic valence dictionary. Thus, it does not select a particular reduced parse, but the corresponding valence frame.[2] As a consequence, there are sentences which have more than one reduced parse corresponding to the chosen frame, but we have no means to decide which of them is correct.

Finally, we provide a list of semantic categories for each semantic head of an NP or a PP. Our goal is to disambiguate these categories.

[1] Such sentences contain no nouns and have no semantic categories of nouns to disambiguate. Thus, they would have a single semantic frame so they would not influence the performance of the algorithm.

[2] The algorithm could be applied for entire parses as well, but then a much bigger set of sentences is needed.

Let us consider the sentence (1) to illustrate the above process with an example. First, the parser *Świgra* produces 7 parses of the sentence, which are compacted into 4 reduced parses presented in (1a). The 4 reduced parses represent actually 3 valence frames, shown (in <> brackets) just before each reduced parse. The EM algorithm selects one of these frames. However, for this sentence the selected frame represents two reduced parses indicated in (2b). The semantic heads of NPs and PPs are then supplied with a list of corresponding categories (cf. table 1).

(1) a.
```
% 'Ona nie wzięła się z twardych reguł wolnego rynku.'
```
(She/It hasn't emerged from hard rules of the free market.)
```
% trees: 7
<wziąć_się :np:nom: :prepnp:z:gen: :sie:>
0-9  wziąć     neg:fin:sg:f:ter
               [0-1:np:on:sg:nom:f:ter,
               1-4:sie,
               4-9:prepnp:z:reguła:gen]
<wziąć_się :np:nom: :prepnp:z:gen: :sie:>
0-9  wziąć     neg:fin:sg:f:ter
               [0-1:np:on:sg:nom:f:ter,
               1-4:sie,
               4-9:prepnp:z:rynek:gen]
<wziąć_się :np:gen: :np:nom: :prepnp:z:gen: :sie:>
0-9  wziąć     neg:fin:sg:f:ter
               [0-1:np:on:sg:nom:f:ter,
               1-4:sie,
               4-7:prepnp:z:reguła:gen,
               7-9:np:rynek:sg:gen:m3:ter]
<wziąć_się :np:gen: :np:nom: :sie:>
0-9  wziąć     neg:fin:sg:f:ter
               [0-1:np:on:sg:nom:f:ter,
               1-4:sie,
               4-9:np:rynek:sg:gen:m3:ter]
```

b.
```
% 'Ona nie wzięła się z twardych reguł wolnego rynku.'
```
(She/It hasn't emerged from hard rules of the free market.)
```
<wziąć :np:nom: :prepnp:z:gen: :sie:>
0-9  wziąć     neg:fin:sg:f:ter::
               [0-1:np:on:sg:nom:f:ter:: pron,
               1-4:sie,
               4-9:prepnp:z:reguła:gen:: UMY]
0-9  wziąć     neg:fin:sg:f:ter::
               [0-1:np:on:sg:nom:f:ter:: pron,
               1-4:sie,
               4-9:prepnp:z:rynek:gen:: msc pos]
```

4. Analysis of the Behaviour of the EM Selection Algorithm

In order to disambiguate categories of head nouns in NPs and PPs in reduced parses of the sentences we process, we have adapted the EM selection algorithm initially used by Dębowski (2007) to select a valence schema of a sentence. Similarly as in the original approach, we have not worked on entire parses, but we have split syntactic-semantic valence frames in such a way that each NP/PP has only one category assigned. The disambiguation process consists in selecting (using the EM algorithm) the most probable frames. We have

considered two versions of the algorithm (Hajnicz to appear): the first treats each frame as a whole (called below EM-whole) and the second is based on the assumption that slots appear in the frame independently (henceforth: EM-indep).

So, after splitting, the sentence from example (1) above receives the following three frames in (2):

(2)
```
<wziąć_się :np:nom:[pron] :prepnp:z:gen:[msc]>
<wziąć_się :np:nom:[pron] :prepnp:z:gen:[pos]>
<wziąć_się :np:nom:[pron] :prepnp:z:gen:[umy]>
```

The correct frame is obviously the third one, since the actual semantic head of PP *z twardych reguł wolnego rynku* is *reguła* 'rule' and the only category of this noun is umy (cognition). If we had this information, there would be no need for disambiguation. Unfortunately, without this knowledge the sentence is quite enigmatic, as neither the verb *emerge* nor the subject *she* puts any constraints on the meaning of the semantic head of the PP (anything can emerge or not from anything). Therefore, there are no external sources of knowledge to determine its semantic category. As a result, the algorithm in both versions was not able to make any decision, as all three frames obtained the same probability. Thus, its failure for this very example is caused by the preprocessing phase.

Before we present an initial evaluation of the algorithm quality, let us emphasise the sources of errors resulting from the preprocessing phase. We should bear in mind that the corpus we use was not manually prepared by experts but it has been automatically annotated by a sequence of programs, each working on the results of the previous one.

4.1 Linguistic Issues

Some kinds of errors are caused by linguistic phenomena themselves. Two such obvious phenomena are homonymy and syncretism. For instance, in sentence (3) the adverb *potem* 'next' was recognised as the instrumental form of the noun *pot* 'sweat'. The indicated syntactic frame has been chosen by the EM algorithm as the most probable one. As the result, there is no way to correctly interpret this sentence semantically.

(3)
```
% 'Zaraz POTEM rozpocznie się uroczysta msza na płycie rynku.'
```
(Just next a ceremonial mess will start on the market place.)
```
<rozpocząć :NP:INST: :np:nom: :sie:>
0-9   rozpocząć   aff:fin:sg:_:ter
                  [0-2:np:POT:SG:INST:m3:ter,
                  2-4:sie,
                  4-9:np:msza:sg:nom:f:ter]
```

Syncretism is usually less problematic, as the lemma of the word is the same, and so are its semantic categories. However, we should remember that in such a case the word (head of a top-most phrase) is in a wrong case, so the phrase is associated with an invalid syntactic slot. Thus, the algorithm puts it into an invalid set of slots (a frame), which influences the results.

Such improper interpretations can result from an incorrect form of the sentence. For instance, in sentence (4) the verb *przejść* 'undergo' has a female gender form, which blocks the (male) noun *rynek* 'market square' to be the subject in nominative case. The sentence is incorrect and has no proper parse. Nevertheless, *Świgra* proposed some (improper) parses for such a sentence. Namely, it considered the noun *rynek* a complement (in accusative) of the verb whose subject has been omitted (subject ellipsis). On the other hand, the noun *remont* 'renovation' was treated as an (accusative) complement of the preposition *w* 'in'. This parse has been selected by the EM algorithm.

(4)
```
% 'Rynek przeszła w tym roku gruntowny remont.'
```
(The market square undergone a major reconditioning this year)
```
<przejść :np:acc: :np:nom: :prepnp:w:acc:>
0-7  przejść    aff:fin:sg:f:ter
                [0-1:np:rynek:sg:acc:m3:ter,
                2-7:prepnp:w:remont:acc]
```

Another error-prone phenomenon are lexicalisations. We do not have any lexicon of idioms, hence they are treated as any other phrases. This also affects the behaviour of the algorithm. Consider the sentence *Strasznie nas na początku robili w konia*. Its literal translation is 'At the beginning [they] awfully made us in(to) a horse'. However, *robić kogoś w konia* is an idiomatic expression meaning 'to take sb for a ride, to con, to swindle him'. Moreover, a similar construction *robić w balona* 'make in(to) a balloon' means almost the same. Evidently, information that *a horse* is an animal and *a balloon* is an artifact has nothing to do with the real meaning of the sentence.

This problem concerns also metaphoric use of words and whole constructs. For instance, the sentence *Problem zaczyna się gdzieś pomiędzy zabudowaniami na Starym Rynku.* 'A/The problem starts somewhere between the buildings on the Old Market Square.' means that the sources of problems can be found somewhere on the Old Market Square. Literary speaking, problems do not *begin between* anything.

The last language-dependent source of errors is free word order of Polish. Consider two sentences (5) and (6). In the first one the verb *rozpocząć* 'begin' separates the NP and the PP whereas in the second one the PP is positioned just after the NP. Consequently, in the second sentence the PP can be considered a modifier of the NP whereas in the first one it cannot. Thus, the list of reduced parses for the second sentence is longer than for the first one. Unfortunately, the EM algorithm tends to choose shorter parses, as it does in this case.

(5)
```
% 'Pierwszy etap rozpocznie się na olkuskim rynku.'
```
(The first stage will start on Olkusz market place.)
```
<rozpocząć :np:nom: :prepnp:na:loc: :sie:>
0-7    rozpocząć    aff:fin:sg:_:ter
                    [0-2:np:etap:sg:nom:m3:ter,
                    2-4:sie,
                    4-7:prepnp:na:rynek:loc]
```

(6)
```
% 'Rozpoczęły się prace modernizacyjne na olkuskim rynku.'
```
(Modernisation works have started on Olkusz market place.)
```
<rozpocząć :np:nom: :sie:>
0-7    rozpocząć    aff:fin:pl:nm1:ter
                    [0-2:sie,
                    2-7:np:praca:pl:nom:f:ter]
```

4.2 Processing Issues

Actually, all steps of the analysis are potential sources of errors. First, sentence boundaries could be improperly delimited. Abbreviations ending with a dot are a typical problem, even more so as we do not have a lexicon of abbreviations. If they do not have another interpretation, the sentence is simply rejected. However, there are abbreviations which can be used as regular words, e.g., *ul.ica* 'str.eet', *im.ienia* 'named', *proc.ent* '%' meaning 'beehive', 'them' and 'catapult' (plural genitive), respectively. Consider the sentence (7). The word *im* has been interpreted as a plural dative of pronoun *on* 'he' forming a false argument of the verb.

(7) `% 'Po niej uczestnicy przejdą w korowodzie przez Rynek`
do Teatru `im.'` [Juliusza Słowackiego ...]
(After that the participants will walk in a procession across the Market place to the [Juliusz Słowacki] Theatre them.)

```
<przejść :np:dat: :np:nom: :prepnp:po:loc: :prepnp:w:loc:>
0-11   przejść   aff:fin:pl:_:ter
                 [0-2:prepnp:po:on:loc,
                 2-3:np:uczestnik:pl:nom:m1:ter,
                 4-10:prepnp:w:korowód:loc,
                 10-11:np:on:pl:dat: :ter]
```

The errors caused by the parser arise from the fact that the syntactic valence dictionary was not ready by the time of processing the corpus,[3] and we used its preliminary version, i.e., the parser was accepting almost any partitioning of a sentence into phrases. This resulted in the proliferation of the parses, many of them inadequate. This in turn affected performance of the EM selection algorithm, due to false statistics. As a result, we can obtain the following improper reduced parses:

- *short:* a separate phrase is absorbed by another one as its modifier (cf. (6));
- *long:* a modifier of another phrase is treated as a separate phrase;
- *wrong:* the two above problems occur or the phrase is improperly interpreted (cf. (3)) or the sentence boundaries are improperly recognised (cf. (7)).

5. Evaluation of the Algorithm

We do not have yet a hand-annotated set of sentences with phrase boundaries and semantic categories assigned by experts. Therefore, we cannot evaluate the quality of the algorithm by usual statistic measures. Nevertheless, we decided to show the properties of the algorithm on a small set of sentences containing a selected noun, *rynek* 'market'. We decided to focus on sentences containing this word because it has two specific, clearly different meanings, which manifests itself by two different categories assigned to it, as shown in table 3. The numbers show how many times the corresponding categories appear in the whole set of sentences before and after the corresponding version of the algorithm was applied. Notice that the category location is six to seven times more frequent than possession and that this proportion is stable.

category name	initial	EM-whole	EM-indep	*rynek* meaning
msc (location):	9 399	7 107	6 194	a market place, a square,
pos (possession):	1 421	1 160	998	a financial, labour etc. resources market.

Table 3: Two semantic categories associated with noun *rynek*

In the set of 41,793 analysed sentences, there are 111 ones containing the word *rynek*. 79 of them have a correct valence frame selected by the EM algorithm, 25 have a short frame selected and seven — a wrong one. The noun *rynek* is present in 50 of the corresponding reduced parses. In ten cases, there are two parses corresponding to a frame: one with the word *rynek* as the semantic head of the phrase and one where *rynek* is not the head (cf. (1) b). For 61 sentences the noun *rynek* is absent in the corresponding parses, 18 of them identified as short ones. For eleven of them the phrase containing the word *rynek* is the one which was

[3] We should stress here that the parsing process of the set of 165,253 sentences in our experiment took one month.

absorbed (cf. (6)). These data are gathered in table 4.

	Proper		Short		Long	Wrong	Sum
rynek present in a reduced parse	OK:	32	OK:	5	0	3	50
	Double:	8:3	Double:	2:1			
rynek absent in a reduced parse		39	OK:	7	0	4	61
			Lack:	11			
							111

Table 4: Properties of the set of sentences containing the word *rynek*

In what follows, we will focus on 47 sentences having a proper (or at least short) valence frame selected with the word *rynek* present in at least one corresponding reduced parse. For these sentences, we have made a manual evaluation of noun categories. Table 5 (a) sums up how many times each category of noun *rynek* has been chosen by an expert or by the algorithm. Column both means that no decision was made; column other means that a parse not containing the noun *rynek* has been chosen.

(a)

	msc	pos	both	other
hand	12	23	8	4
indep	36	5	2	4
whole	29	2	13	3

(b)

msc			pos			both			other	
=	≠	⊂	=	≠	⊂	=	≠	⊃	=	≠
12	20	4	3	0	2	1	0	1	3	1
9	14	5	2	0	0	2	0	11	1	3

Table 5: Two categories of *rynek* in numbers

The presented values are definitely surprising. First, hand-made annotation shows that the word *rynek* is two times more often used in the context of possession than in the context of location, contrary to the general proportion. In contrast, both versions of the algorithm tend to prefer interpreting the noun as a location even more often than for other nouns. Hence, this is not surprising that evaluation made for this set of sentences is extremely bad. The results are presented in table 5 (b). The EM-indep algorithm makes 25 correct decisions (we assume selecting one of the two categories assigned by an expert as a correct decision) and 22 incorrect decisions, which means 53.2% of correct decisions. The EM-whole algorithm makes 19 correct decisions, 17 incorrect decisions, whereas 11 cases are left undecided.

Below we investigate the reasons for such poor behaviour of the algorithm. The first, probably the most important reason is sparseness of the data. Recall that we consider two kinds of syntactic slots: NPs which can appear in five cases and PPs which may appear in a form of 60 pairs <pronoun, case>. Then, we have 26 semantic categories at our disposal, from one to six categories per noun. Obviously, only a subset of this repertoire of slots appears with particular verbs. Our future work encompasses evaluation of the algorithm results w.r.t. verbs, but it seems that it would be directly proportional to the number of sentences containing the verb and inversely proportional to the number of valence frames of the verb. Notice that we have from one to four PPs/NPs per sentence, hence the data sparseness for EM-whole algorithm is even more important. This is an explanation why the algorithm leaves more cases undecided (cf. also table 3).

Linguistic phenomena and preprocessing issues discussed in section 3 affect results as well. Let us discuss yet another example. Sentence (8) has two reduced parses, in which *rynek* is once the subject and once the object.[4] As a result, we have eight syntactic-semantic valence

[4] In all examples, we mark proper categories by small capitals and underline the ones chosen by the algorithm.

frames instead of four (for one syntactic frame!). Both versions of the algorithm have chosen the same, wrong frame (<dzielić :acc:[msc] :nom:[wytw]>).

(8)
```
% 'Kolej dzieli rynek.' ([The] railway (company) divides the market.)
<dzielić :np:acc:[czas wytw msc pos] :np:nom:[czas wytw msc pos]>
0-3   dzielić   aff:fin:sg:_:ter::
                [0-1:np:kolej:sg:nom:f:ter:: WYTW czas,
                2-3:np:rynek:sg:acc:m3:ter:: msc  POS]
0-3   dzielić   aff:fin:sg:_:ter::
                [0-1:np:kolej:sg:acc:f:ter:: wytw czas,
                2-3:np:rynek:sg:nom:m3:ter:: msc pos]
```

The above remarks concern the overall results. Now we want to focus on sentences containing the word *rynek*. Our first observation is that the noun tends to appear in the subject position with verbs having an animate subject, such as *lubić* 'to like', *cieszyć się* 'to enjoy', *powiedzieć* 'to tell'; the verb *dzielić* 'to divide' from the previous example can be included in the same group. The statistics of the initial data (before disambiguation) for verb *lubić* is presented in table 6. As you can see, a typical subject of *lubić is* a person, and a place appears on that position more often than possession. Next, an object of *lubić* is usually an artifact, an act or a place (cf. table 1). Thus, it is obvious that such atypical use of the verb as in sentence (9) affects the behaviour of the algorithm, as there is a small amount of data supporting correct frames, hence we obtain random results. Observe that for a similar sentence *Rynek lubi słońce.* '[The] market likes the sunshine.' we would prefer to assign the category location to *rynek*.

lubić		487								
nom	all:	385	os:	339	msc:	5	pos:	1		
acc	all:	447	wytw:	60	czy:	58	msc:	44	zw:	14

Table 6: The verb *lubić* in numbers

(9)
```
% 'Rynek lubi fuzje.' (The market likes fusions.)
<lubić :np:acc:[msc pos] :np:nom:[wytw zw]>
0-3   lubić   aff:fin:sg:_:ter::
              [0-1:np:rynek:sg:nom:m3:ter:: msc POS,
              2-3:np:fuzja:pl:acc:f:ter:: wytw ZW]
```

Next, consider sentence (10). People usually escape from places (location) rather than from abstract objects denoted as possession, hence the choice of the category location is straightforward. The only suggestion that the proper choice is possession is that we talk about *labour market*. Unfortunately, in our approach, modifiers are ignored. Nevertheless, one category of the noun *praca* 'work' is location, hence the construction *rynek pracy* is similar to *rynek miasta* 'town's market', so this information would not help. However, similar modifiers (cf. *Ciekawe rzeczy dzieją się na światowym rynku surowców.* 'Interesting things happen on the world resource market.' or *Na rynku funduszy inwestycyjnych trudno mówić o dobrym tygodniu,* 'It is hard to talk about a good week on the market of investment funds.') should be useful, as places are usually not modified by possession (the semantic category of *fund*). Consider also a similar sentence *Inwestorzy gremialnie uciekają z naszego rynku.* 'Investors as one man escape from our market'. The subject *inwestorzy* suggests a financial market. However, *inwestor* is categorised simply as a person, so no specific information is available. Finally, observe that sentence (10) has two reduced parses (an improper one signed with x). Since the category location was chosen for prepnp:z:gen, then both parses were accepted, as

this is the category of *rynek* as well as *praca*. The reason is that the algorithm operates on syntactic-semantic valence frames, not on reduced parses themselves. Thus, this phenomenon is the feature of the method (or rather of the fact that we obtain more than one reduced parse per sentence), and cannot be treated as an error in cases when a semantic category itself is chosen correctly for a noun in a correct parse.

(10) % 'Uciekli z rynku pracy.' *([They] escaped from the labour market.)*
<uciec :np:nom:[pron] :prepnp:z:gen:[msc pos czy wytv]>
0-4 uciec aff:fin:pl:m1:ter
[1-4:prepnp:z:rynek:gen:: msc POS]
0-4 uciec aff:fin:pl:m1:ter
[1-4:prepnp:z:praca:gen:: czy wytw msc] **X**

6. Conclusions and Future Work

In this paper we presented results of our first experiments in word sense disambiguation of heads of top-most phrases in a sentence. After the initial evaluation, made on an arbitrarily chosen set of sentences, the proposed method seems poor. However, actually we need to apply some other methods to the same data. Then we will be able to compare the results and estimate the real quality of our approach. We think that it is promising, even though it needs revisions.

Thus, we have focused on analysing potential reasons for such unsatisfactory results. Some of them are intrinsic properties of our data (the analysed corpus). The others show us a way for improvements.

First, we should reduce the problem of data sparseness. The algorithm should be tested on a bigger set of sentences per verb with a smaller set of valence frames. We want to obtain this by applying more restrictive version of the *Świgra* parser. Next, we want to operate on a more fine-grained wordnet-like information about noun senses. In the future, we want also to incorporate information coming from modifiers of a noun. However, these two modifications entail increase of sparseness of algorithm's data.

We should stress that the actual evaluation of the algorithm is probably much better as many sentences have single reduced parses with the univocal semantic category of each noun appearing in it, hence their interpretation has to be correct. Sentences containing the two senses of word *rynek* evidently do not have such parses.

To conclude, it should be emphasized that our goal is creation of a semantic valence dictionary, not WSD itself. Observe that the algorithm tends to select most frequent categories, sometimes contrary to the real meaning of a noun in a given context. However, our observations show that it prefers syntactic-semantic valence frames that are valid when the context is slightly different. Thus, we obtain irrelevant frames relatively rarely (cf. (9)). Such "noisy" frames should be cast off during pruning phase of dictionary creation.

It is rather simple to collect syntactic-semantic valence frames selected by the algorithm discussed in the paper. However, creating a separate dictionary entry for every tuple of pairs <syntactic slot, category> possible for a particular verb would cause proliferation of entries. Moreover, most of these entries would describe the verb with the same meaning. For instance, we would have a set of entries for *kupić* 'to buy', cf. (11):

(11) <:np:nom: [os] :np:acc:[jedz]>, <:np:nom: [os] :np:acc:[msc]>,
<:np:nom: [os] :np:acc:[rsl]>, <:np:nom: [os] :np:acc:[zwz]>,
<:np:nom: [os] :np:acc:[sbst]>, <:np:nom: [os] :np:acc:[wytw]>,

instead of one:

(12) <:np:nom: [os] :np:acc:[jedz msc rsl zwz sbst wytw]>.

Finding a method to aggregate entries that groups those with the same meaning and separate those with different meanings (i.e., detecting polysemy) is an important and not trivial task.

References

Abney, S. (2004) Understanding the Yarowsky algorithm. *Computational Linguistics*, 30(3): 365–395.

Agirre, E. and P. Edmonds, eds (2006) *Word Sense Disambiguation. Algorithms and Applications*. Dordrecht: Springer-Verlag.

Baker, C. F., C. J. Fillmore and J. B. Lowe (1998) The Berkeley FrameNet Project. In *Proceedings of the COLING-ACL'98 Conference*, pages 86–90, Montreal, Canada.

Banerjee, S. and T. Pedersen (2002) An adapted Lesk algorithm for word sense disambiguation using WordNet. In A. F. Gelbukh, ed, *Proceedings of the 3rd Conference on Computational Linguistics and Intelligent Text Processing CICLing-2002*, 136–145, Mexico City. Springer-Verlag.

Boas, H. C. (2002) Bilingual FrameNet dictionaries for machine translation. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, 1364–1371, Las Palmas: ELRA - European Language Ressources Association.

COLING (2002) *Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002)*, New Brunswick: Morgan Kaufmann.

Dębowski, Ł. (2007) Valence extraction using the EM selection and co-occurrence matrices. arXiv. Available at: http://arxiv.org/ps_cache/arxiv/pdf/0711/0711.4475v1.pdf.

Dębowski, Ł. and M. Woliński (2007) Argument co-occurrence matrix as a description of verb valence. In Z. Vetulani, editor, *Proceedings of the 3rd Language & Technology Conference*, pages 260–264. Poznań: Wydawnictwo Poznańskie.

Derwojedowa, M., M. Piasecki, S. Szpakowicz, and M. Zawisławska (2007) Polish WordNet on a shoestring. In *Data Structures for Linguistic Resources and Applications: Proceedings of the GLDV 2007 Biannual Conference of the Society for Computational Linguistics and Language Technology*, Tübingen: Universität Tübingen.

Dorr, B. J. and D. Jones (1996) Role of word sense disambiguation in lexical acquisition: Predicting semantics from syntactic cues. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-1996)*, 322–327, Copenhagen: Morgan Kaufmann.

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

Gaustad, T. (2004). Linguistic knowledge and word sense disambiguation. PhD thesis, Groningen: Rijksuniversiteit.

Hajnicz, E. (2007) *Dobór czasowników do badań przy tworzeniu słownika semantycznego czasowników polskich*. Technical Report 1003, Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Hajnicz, E. (to appear) Semantic annotation of verb arguments in shallow parsed Polish sentences by means of EM selection algorithm.

Hajnicz, E. and A. Kupść (2001) *Przegląd analizatorów morfologicznych dla języka polskiego*. Technical Report 937, Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Ion, R. and D. Tufiş. (2004) Multilingual word sense disambiguation using aligned wordnets. *Romanian Journal of Information Science and Technology*, 7(1–2): 183–200.

Král, R. (2001) Three approaches to word sense disambiguation for Czech. In *Proceedings of the 4th International Conference on Text, Speech and Dialogue (TSD-2001)*, 174–179, Berlin: Springer-Verlag.

Levin, B. (1993) *English Verb Classes and Alternation: A Preliminary Investigation*. Chicago: University of Chicago.

Lin, D. and P. Pantel (2002) Concept discovery from texts. In COLING (2002), 577–583.

McCarthy, D. and J. Carroll (2003) Disambiguating nouns, verbs and adjectives using automatically acquired selectional preferences. *Computational Linguistics*, 29(4): 639–654.

Mędak, S. (2005) *Praktyczny Słownik Łączliwości Składniowej Czasowników Polskich*. Cracow: Universitas.

Obrębski, T. (2002) Automatyczna analiza składniowa języka polskiego z wykorzystaniem gramatyki zależnościowej. PhD thesis, Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Przepiórkowski, A. (2006) What to acquire from corpora in automatic valence acquisition. In V. Koseska-Toszewa and R. Roszko, eds, *Semantyka a Konfrontacja Językowa*, volume 3, Warsaw: Slawistyczny Ośrodek Wydawniczy.

Przepiórkowski, A. (2007a) On heads and coordination in valence acquisition. In A. Gelbukh, ed, *Proceedings of the 8th Conference on Computational Linguistics and Inteligent Text Processing CICLing-2007*, 50–61, Mexico City: Springer-Verlag.

Przepiórkowski, A. (2007b) A preliminary formalism for simultaneous rule-based tagging and partial parsing. In G. Rehm, A. Witt and L. Lemnitzer, eds, *Data Structures for Linguistic Resources and Applications: Proceedings of the Biennial GLDV Conference*, 81–90. Tübingen: Gunter Narr Verlag.

Przepiórkowski, A., A. Kupść, M. Marciniak and A. Mykowiecka (2002) *Formalny Opis Języka Polskiego. Teoria i Implementacja*. Warsaw: Akademicka Oficyna Wydawnicza Exit.

Rabiega-Wiśniewska, J. (2004) Podstawy lingwistyczne automatycznego analizatora morfologicznego Amor. *Poradnik Językowy* 10: 59-78.

Schütze, H. (1998) Automatic word sense discrimination. *Computational Linguistics*, 24(1): 97–123.

Suárez, A. and M. Palomar (2002) A maximum entropy-based word sense disambiguation system. In COLING (2002), 960–966.

Świdziński, M. (1994) *Syntactic Dictionary of Polish Verbs*. Uniwersytet Warszawski / Universiteit van Amsterdam.

Vossen, P., ed (1998) *EuroWordNet: a Multilingual Database with Lexical Semantic Network*. Dordrecht: Kluwer Academic Publishers.

Woliński, M. (2004) Komputerowa weryfikacja gramatyki Świdzińskiego. PhD thesis, Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Woliński, M. (2006) Morfeusz — a practical tool for the morphological analysis of Polish. In M.A. Kłopotek, S. T. Wierzchoń and K. Trojanowski, eds, *Proceedings of the Intelligent Information Systems New Trends in Intelligent Information Processing and Web Mining IIS:IIPWM'06*, Advances in Soft Computing, 503–512, Ustroń: Springer-Verlag.

Zawisławska, M. and M. Derwojedowa, M. (2008) Opis rzeczowników w ramach sieci semantycznej typu wordnet. submitted to *Semantyka a Konfrontacja Językowa*.