Intelligent Information Systems 9999
ISBN 666-666-666, pages 1–14

How Valence Information Influences Parsing Polish with *Świgra*

Elżbieta Hajnicz and Marcin Woliński

Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

Abstract

In this paper we try to establish how information about valence of Polish verbs influences parsing. The experiments were performed with the *Świgra* parser used with and without a valence dictionary. The results concerning the number of accepted sentences, the number of parse trees and reduced parses per sentence are presented. We also analyse how the availability of valence information influences syntactic-semantic annotation of parsed corpus.

1 Introduction

In this paper we characterise the process of syntactic analysis of Polish sentences performed by the *Świgra* parser. *Świgra* was implemented by Woliński (2004) on the basis of a metamorphosis grammar of Polish GFJP created by Świdziński (1992). This version of *Świgra* was operating with a small valence dictionary composed of about 300 verbs.

More recently, the parser was applied to creation of syntactic valence dictionary of Polish verbs (Dębowski, 2007; Dębowski and Woliński, 2007). The parser was adapted to work without any valence dictionary, i.e., its "permissive" version accepted any combination of arguments for any verb, and then the results were filtered with statistical methods.

The current paper is in a way a by-product of a project where parsing a corpus with *Świgra* was the first step of the process of semantic annotation of verb arguments in Polish sentences (cf. Hajnicz, 2009). *Świgra* was used with a handcrafted valence dictionary (Świdziński, 1994) supplemented with frames obtained automatically by Dębowski (2007). Therefore, this work is actually the first attempt to use the parser with an extensive valence dictionary.

In the paper we investigate the influence of valence information on parsing. We compare the results of *Świgra* run with and without a valence dictionary and we provide some statistics counted on experimental data.

2 The Corpus

Our source of text is the IPI PAN Corpus of written Polish (Przepiórkowski, 2004), referred to as Kipi. The texts are segmented into paragraphs and sentences and

annotated with morphosyntactic tags. The 2nd edition of the corpus contains 250 mln words.

For the purpose of semantic annotation (Hajnicz, 2009), we have selected a small subcorpus of 195 042 sentences, referred to as SemKipi. The selected sentences contain:

(a) only words that can be interpreted by the morphosyntactic analyser *Morfeusz* (Woliński, 2006), which is used by *Świgra*,
(b) one or more verbs from the set of 32 verbs chosen manually according to the rules described by Hajnicz (2007),
(c) at most three verbs in all.

SemKipi contains a total of 5472 verbs. However, according to Zipf law, their frequency decreases rapidly.[1] The frequency of as many as 1600 verbs is 1.

3 The Valence Dictionary

We have rejected the idea of using the dictionary automatically generated by Dębowski (2007) for the process of semantic annotation and so for the experiments presented here. This dictionary has several versions. Unfortunately, the version generated under the most restrictive values of some parameters contains only rather simple frames to satisfy our needs, while the less restrictive versions contain too many erroneous frames.

Consequently, an extensive valence dictionary was prepared specially for our task. Its main component is Świdziński's (1994) valence dictionary, which contains 1064 verbs.[2] Since many frequent verbs are absent in this dictionary, we had to extend it with frames from other sources.

First of all, we checked the dictionary against 32 verbs of the benchmark set given in (Hajnicz, 2007). Only 24 of them were included in Świdziński's dictionary. We managed to adjust the available entries of aspectual counterparts of further 4 verbs.[3] Entries for the remaining 4 verbs were elaborated by analysing entries of the valence dictionary automatically created by Dębowski (2007). All entries of this part of the dictionary were carefully studied, modified and augmented.

Next, similar additions were done for other verbs. We adjusted the entries of the missing 269 aspectual counterparts of verbs described in Świdziński's dictionary. We used also the restrictive version of the valence dictionary extracted by Dębowski (2007), from which the entries of 955 verbs with frequency greater than 5 were added verbatim to increase the coverage of our dictionary on SemKipi.

Apart from extending the dictionary, we deleted 135 verbs, which frequency results from homonymy, such as *meczeć* (1729, *to bleat*), which imperative form is a homonym of the noun *mecz* (*match*).

We identified 950 verbs in SemKipi (apart from 24 verbs mentioned above) in Świdziński's dictionary, including 220 with frequency lower than 6. Other 3170

[1] In this set of sentences, the Zipf law does not cover the manually chosen verbs, because of the way of choosing sentences.

[2] If we consider verbs with *się* (reflexive marker) as different entities, the number of verbs increases to 1379.

[3] In Polish, aspect is accomplished by different verbs.

such rare verbs were ignored. Unfortunately, neither Świdziński's nor the automatically obtained dictionary cover as many as 80 frequent verbs. We did not prepare entries for them. The whole resultant dictionary contains about 2290 verbs.

The maximal number of syntactic frames per verb in the whole dictionary is 25, their mean is 2.8 and their median is 2. The maximal number of arguments per frame (including subject and reflexive marker *się*) is 5, the mean is 2.5 and the median is 3.

3.1 Frame structure

In this section we present the format used to represent valence frames in *Świgra*. Frames can be interpreted as lists of arguments, two of them having a special status, namely nominative subject NP and the reflexive marker *się*. Verbs which allow for a subject are called *proper* and denoted with 'V', whereas verbs without nomitative subject are called *quasi-verbs* or *improper* and denoted with 'Q'. Every entry in the dictionary is composed of three parts separated with tabulation:

1. a lemma, possibly with the reflexive marker *się*,
2. the letter 'V' for proper verbs or 'Q' for quasi-verbs,
3. a list of other arguments separated with '+' signs.

The list of arguments can include:

1. adjectival phrases AdjP,
2. adverbial phrases AdvP,
3. infinitive phrases InfP,
4. nominal phrases NP,
5. prepositional-adjectival phrases PrepAdjP,
6. prepositional-nominal phrases PrepNP,
7. clauses SentP.

Some arguments are parametrised. The only parameter of AdjP and NP is their case, the only parameter of InfP is its aspect. PrepAdjP and PrepNP have two parameters: a preposition and the case of its AdjP or NP complement, respectively. SentP has one parameter, namely the complementizer introducing the clause. *Oratio recta* was ignored, since it is not covered by GFJP.

A special case of arguments containing clauses are so-called *correlats*. In such constructions a clause follows an NP or PrepNP, in which NP is just composed of the single pronoun *to* (*this*). They are represented as a special case of SentP argument with two or three parameters: optional preposition, the case of *to* and complementizer.

Below we list examples of dictionary entries for verbs *lubić* (*like*), *odnieść* (*carry back, achieve*), *odnieść się* (*treat, concern*) and *znać* (*know*).

(1)	lubić	V	infp(_)	
	lubić	V	np(acc)	—
	lubić	V	np(acc)+prepnp(‘za’, acc)	
	lubić	V	sentp(‘by’)	
	lubić	V	sentp(‘jak’)	
	lubić	V	sentp(‘kiedy’)	
	lubić się	V	prepnp(‘z’, inst)	
(2)	odnieść	V	np(acc)	—
	odnieść	V	np(acc)+advp	
	odnieść	V	np(acc)+np(dat)	
	odnieść	V	np(acc)+prepnp(‘do’, gen)	
	odnieść się	V	prepnp(‘do’, gen)+prepnp(‘w sprawie’, gen)	
	odnieść się	V	prepnp(‘z’, inst)+prepnp(‘do’, gen)	
(3)	znać	Q	sentp(‘że’)	
	znać	Q	np(acc)+prepnp(‘po’, loc)	—
	znać	Q	np(acc)+prepnp(‘po’, loc)+sentp(‘że’)	
	znać	V	np(acc)	—
	znać	V	np(acc)+prepnp(‘z’, gen)	
	znać się	V	prepnp(‘na’, loc)	
	znać się	V	prepnp(‘z’, inst)	

3.2 The Problem of Subframes

Świdziński's dictionary includes frames that are subframes of other frames. The idea was to list all non-elliptic frames. For instance, one of possible frames for the verb *lubić* is `V np(acc)` (cf. 1), which is instantiated in the sentence *Ktoś nie lubi dyrektora?* (*Does anyone not like the manager?*, cf. the example 6 below). But the dictionary contains as well a larger frame `np(acc)+prepnp('za',acc)` (cf. sentence *Ktoś nie lubi dyrektora za bezwzględność?*, *Does anyone not like the manager for his ruthlessness?*.) However, *Świgra* accepts already all subframes of each frame listed in the dictionary, to automatically account for the phenomenon of ellipsis. Hence, listing subframes in the valence dictionary only slows down parsing. Thus, we deleted automatically all subframes from the dictionary. All non-elliptic frames being subframes of other frames were marked in the above examples with a −. Please note, however, that special arguments, i.e., nominative subject and reflexive marker, did not undergo this procedure.

4 Improvements in the *Świgra* parser

GFJP and consequently *Świgra* have some known deficiencies, which diminish the corpus coverage of the parser (cf. Woliński, 2004, 2005). Fortunately, these do not limit the types of verb arguments possible, so should not have much impact on the process of (semantic) valence extraction. We have made, however, some improvements to the parser with respect to the reflexive marker *się*.

The original version of *Świgra* assumed that the reflexive marker is a part of a verb form. As a consequence, only sentences with *się* positioned just after or before

a verb were parsed. This limitation was overcome by treating *się* as a separate argument of the verb. The only difference is the fact that it cannot be ellipted.

Polish has impersonal forms of verbs in past tense. In present tense, impersonal statements are expressed by constructions with neuter singular 3rd person verb form supplemented with *się*, for example

(4) *Ostatnio dużo mówi się o kryzysie*
Recently a lot speak refl. about crisis
'Recently [people] talk a lot about the crisis.'

In order to parse such sentences, we added a nominative neuter plural pronoun interpretation for *się* (which means *się* gets interpreted as the subject of such sentences).

5 Reduced parses

Świgra produces forests of complete parse trees. However, for semantic verb argument annotation we do not need complete trees. We are only interested in verbs and their arguments. For this reason, we reduce parses to their shallow form, which we call *reduced parses*. These can be seen as instantiated valence frames. For each argument its type, head and its morphological characteristics are given.

Lists of reduced parses obtained by "permissive" *Świgra* are presented in (5), (6), and (7). Number of trees produced with the use of valence dictionary are given in brackets. The reduced parses obtained from them are marked with a +.

(5) `% Giełda lubi płatać figle.`
stock market likes to play tricks
'The stock market likes to play tricks on us.'

```
% trees:  7  (3)
0-4  lubić aff:fin:sg:_:ter                                    +
     [0-1:np:giełda:sg:nom:f:ter, 2-3:infp:płatać:imperf]
0-4  lubić aff:fin:sg:_:ter  [0-1:np:giełda:sg:nom:f:ter,
     2-3:infp:płatać:imperf, 3-4:np:figiel:pl:acc:m3:ter]
0-4  lubić aff:fin:sg:_:ter                                    +
     [0-1:np:giełda:sg:nom:f:ter, 2-4:infp:płatać:imperf]
0-4  płatać aff:inf:ter      [0-2:adjp:luby:pl:nom:m1]
0-4  płatać aff:inf:ter
     [0-2:adjp:luby:pl:nom:m1, 3-4:np:figiel:pl:acc:m3:ter]
```

(6) `% Ktoś nie lubi dyrektora?`
somebody not likes manager
'Does anyone not like the manager?'

```
% trees:  8  (6)
0-4  lubić neg:fin:sg:_:ter  [0-1:np:ktoś:sg:nom:m1:ter]        +
0-4  lubić neg:fin:sg:_:ter                                    +
     [0-1:np:ktoś:sg:nom:m1:ter, 3-4:np:dyrektor:sg:acc:m1:ter]
0-4  lubić neg:fin:sg:_:ter
     [0-1:np:ktoś:sg:nom:m1:ter, 3-4:np:dyrektor:sg:gen:m1:ter]
```

(7) `% Odniosłem kontuzję kolana.`
sustained-I `injury knee`
'I sustained an injury of my knee.'

```
% trees:  9 (7)
0-4  odnieść aff:fin:sg:m:pri  []                                          +
0-4  odnieść aff:fin:sg:m:pri  [2-3:np:kontuzja:sg:acc:f:ter]              +
0-4  odnieść aff:fin:sg:m:pri
     [2-3:np:kontuzja:sg:acc:f:ter, 3-4:np:kolano:sg:gen:n:ter]
0-4  odnieść aff:fin:sg:m:pri  [2-4:np:kontuzja:sg:acc:f:ter]              +
0-4  odnieść aff:fin:sg:m:pri  [3-4:np:kolano:pl:acc:n:ter]                +
0-4  odnieść aff:fin:sg:m:pri  [3-4:np:kolano:sg:gen:n:ter]
```

The corresponding valence frame for (5) is lubić V infp(_), for (6) the frame lubić V np(acc) corresponds to the second marked reduced parse, whereas the first one was obtained using its subframe lubić V. As for (7), the last three marked reduced parses correspond to the frame odnieść V np(acc), whereas the first one was obtained using its subframe odnieść V. Observe that this time the 1st person subject *I* is elliptic and it is absent in the reduced parses.

6 The Experiments

In this section we compare properties of the two versions of the parser. The first variant is *Świgra* working with the valence dictionary described in section 3, which will be referred to as *d-Świgra*. The "permissive" version of *Świgra* working without a valence dictionary will be refered to as *p-Świgra*. It assumes that every frame is possible for every verb. Frames are limited to at most five arguments and we do not allow for more than one argument of any type (nominal and adjectival phrases in different cases are considered different types).

For the sake of comparison, we randomly selected 10 sets of sentences from SemKipi, each containing 5000 sentences. Each of the sets was parsed with *d-Świgra* and *p-Świgra*.

In the following, we compare the numbers of accepted and rejected sentences; numbers of generated trees and reduced parses. Then the influence of the valence dictionary on the process of semantic annotation is studied.

6.1 Numbers of parsed sentences

We start with comparing the percentage of sentences parsed by each version of the algorithm. There are three possible reasons for which a sentence cannot be parsed by *Świgra*:

- It is not covered by GFJP;
- It contains segments that cannot be interpreted morphologically by Morfeusz;
- Its parsing time exceeded a preselected limit.

The second reason has been eliminated for the test set by careful selection of sentences (cf. section 2). We use a time limit in the parsing, since this is the simplest way to cope with some rare sentences, which have very long parsing time.

In Table 1 we present numbers of sentences accepted (column *accepted*), rejected (column *rejected*), and killed after time limit (column *killed*) by *d-Świgra*. In Table 2 similar data is presented for *p-Świgra*.

	accepted				error		killed		rejected	
	nr.		%		nr.	%	nr.	%	nr.	%
SET 0	2031	(81)	40.6	(1.6)	547	10.9	10	0.20	2412	48.2
SET 1	1981	(88)	39.6	(1.8)	559	11.2	11	0.22	2449	49.0
SET 2	1994	(84)	39.9	(1.7)	593	11.9	7	0.14	2406	48.1
SET 3	1966	(83)	39.3	(1.7)	566	11.3	5	0.10	2463	49.2
SET 4	1924	(83)	38.5	(1.7)	538	10.8	5	0.10	2531	50.6
SET 5	2048	(80)	40.9	(1.6)	523	10.5	10	0.20	2419	48.3
SET 6	2003	(95)	40.1	(1.9)	542	10.8	8	0.16	2447	48.9
SET 7	1972	(81)	39.4	(1.6)	543	10.9	5	0.10	2480	49.6
SET 8	1973	(82)	39.4	(1.6)	597	11.9	9	0.18	2421	48.4
SET 9	2026	(96)	40.5	(1.9)	555	11.1	4	0.08	2415	48.3
TOTAL	19919	(853)	39.8	(1.7)	5563	11.1	74	0.15	24443	48.9
MEAN	1992	(85)	39.8	(1.7)	556	11.1	7	0.15	2444	48.9
DEV	28.70	(4.62)			17.96		2.80		29.90	

	accepted		rejected		killed	
	nr.	%	nr.	%	nr.	%
SET 0	2031	40.6	2959	59.1	10	0.2
SET 1	1981	39.6	3008	60.2	11	0.22
SET 2	1994	39.9	2999	60.0	7	0.14
SET 3	1966	39.3	3029	60.5	5	0.1
SET 4	1925	38.5	3070	61.4	5	0.1
SET 5	2048	40.9	2942	58.8	10	0.2
SET 6	2003	40.1	2989	59.7	8	0.16
SET 7	1972	39.4	3023	60.5	5	0.1
SET 8	1973	39.4	3018	60.3	9	0.18
SET 9	2026	40.5	2970	59.4	4	0.08
MEAN	1992	39.8	3001	60.0	7	0.15
DEV	28.7		29.9		2.8	

TABLE 1: Parsabilty of sentences by means of *d-Świgra*

The last row in both tables (row *DEV*) shows standard deviations of the respective values across the set of 10 experiments. Their low values suggest that the results do not depend on selection of the set of sentences.

In Table 3 we compare the sets of sentences accepted by each version of the parser, showing the number of sentences parsed by *d-Świgra* and not parsed by *p-Świgra* (the 1st column) and *vice versa* (the 2nd column), parsed by both parsers (the 3rd column) and by none of them (the 4th column).

	accepted		rejected		killed	
	nr.	%	nr.	%	nr.	%
SET 0	2175	43.5	2657	53.1	168	3.4
SET 1	2091	41.8	2736	54.7	173	3.5
SET 2	2095	41.9	2746	55	159	3.2
SET 3	2109	42.2	2751	55	140	2.8
SET 4	2067	41.3	2773	55.5	160	3.2
SET 5	2165	43.3	2660	53.2	175	3.5
SET 6	2134	42.7	2730	54.6	136	2.7
SET 7	2136	42.7	2734	54.7	130	2.6
SET 8	2140	42.8	2712	54.2	148	3
SET 9	2184	43.7	2685	53.7	131	2.6
MEAN	2130	42.6	2718	54.4	152	3
DEV	31.28		33.1		15	

Table 2: Parsabilty of sentences by means of *p-Świgra*

	dict−perm		perm−dict		perm∩dict		neither	
	nr.	%	nr.	%	nr.	%	nr.	%
SET 0	95	1.9	240	4.8	1935	38.7	2730	54.6
SET 1	96	1.9	207	4.1	1884	37.7	2813	56.3
SET 2	112	2.2	214	4.3	1881	37.6	2793	55.9
SET 3	98	2.0	241	4.8	1868	37.4	2793	55.9
SET 4	92	1.8	234	4.7	1833	36.7	2841	56.8
SET 5	109	2.2	228	4.6	1937	38.7	2726	54.5
SET 6	95	1.9	226	4.5	1908	38.2	2771	55.4
SET 7	74	1.5	240	4.8	1896	37.9	2790	55.8
SET 8	88	1.8	259	5.2	1881	37.6	2772	55.4
SET 9	92	1.8	251	5.0	1933	38.7	2724	54.5
TOTAL	951	1.9	2340	4.7	18956	37.9	27753	55.5
MEAN	95	1.9	234	4.7	1895	37.9	2775	55.5
DEV	6.92		12.20		26.20		30.70	

Table 3: The comparison of parsabilty of sentences

d-Świgra accepts 39.8% sentences on average, whereas *p-Świgra* accepts 42.6% sentences on average, i.e., 2.8 percentage points more. However, as many as 1.9% of sentences were parsed by *d-Świgra* and were not parsed by *p-Świgra*, all of them exceeded the time limit. They would be accepted if the parser was given enough time, so this suggests that the dictionary indeed guides the parser allowing it to generate results quicker. On the other hand, 4.7% of sentences were parsed by *p-Świgra* and were not parsed by *d-Świgra*, all of them being not accepted. Thus, 5.9% of source sentences were parsed by one of the parsers only.

There are few reasons for rejecting a sentence by *d-Świgra*. First, a sentence itself may be grammatically incorrect. Second, *Świgra* does not accept some correct sentences. For the above cases, *p-Świgra* generates improper parse trees that were inconsistent with valence frames of verbs predicating the clauses in the sentence. Next, one of the verbs in the sentence or its particular frame could be absent from the valence dictionary. For such cases, there exists a proper parse tree among trees produced by *p-Świgra*.

6.2 Numbers of trees and reduced parses

Świgra tends to produce large parse forests. The number of reduced parses of a sentence is much smaller than the number of entire parse trees, the more so as we have only considered actual arguments of verbs (without adjuncts). Nevertheless, this number is still considerable.

In order to fairly compare results of parsing, we decided to count parse trees and reduced parses obtained by each version of parser only for sentences parsed by both of them.

		SET 0	SET 1	SET 2	SET 3	SET 4	SET 5	SET 6	SET 7	SET 8	SET 9	total
parse	d-Świgra	36	32	40	36	40	33	36	36	36	36	36
trees	p-Świgra	208	162	222	166	216	156	176	200	186	172	186
reduced	d-Świgra	7	6	7	7	7	7	7	7	7	7	7
parses	p-Świgra	33	30	36	31	35	30	31	32	33	31	32

Table 4: Medians of numbers of parse trees and reduced parses

In Table 4 we present the medians of numbers of trees and reduced parses per sentence. The differences between the two versions of the parser are striking. The median of trees produced by *p-Świgra* is 5 times greater than the median of trees produced by *d-Świgra*. The proportion of medians for reduced parses is about 4.5.

The distribution of parses is concentrated in small values. However, there are some outliers with very high numbers of parses. In Fig. 1 we present percentiles of numbers of parses for both parsers in their full and reduced variants. Less than 2% of sentences have numbers of trees several orders of magnitude larger than all the other sentences. This phenomenon is especially pronounced in parse trees.

In Table 5 we show how the usage of a valence dictionary influences the number of arguments per verb (counted in reduced parses). Remembering that the number of arguments of a verb in *p-Świgra* was limited to 5 and that *Świgra* considers all subframes (and hence most "subparses") we can deduce that for most of sentences the maximal number of arguments is used. Thus, without this limitation the number of phrases per reduced parse would be probably much larger. The number of parse trees and reduced parses would increase as well.

As one may expect, noun and prepositional phrases are most common arguments. Pay attention to the fact that elipsis of nominal subject (especially for 1st

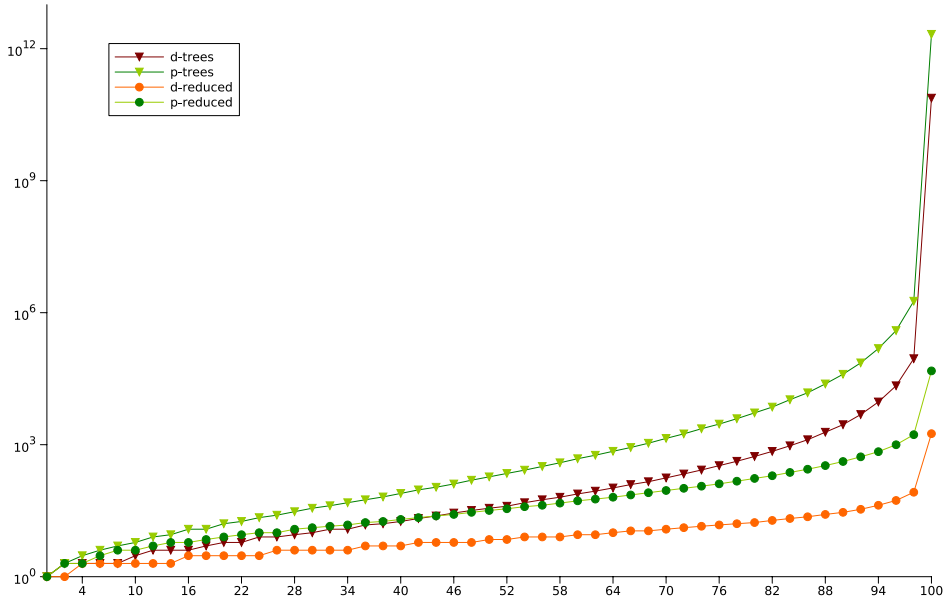

FIGURE 1: Percentiles of numbers of parse trees for *d-Świgra* — d-trees, *p-Świgra* — p-trees, numbers of reduced parses for *d-Świgra* — d-reduced, and *p-Świgra* — p-reduced

and 2nd person verbs) is charactersistic for Polish, which lessens the number of explicit arguments.

6.3 Results of the EM reduced parse selection

Before applying semantic annotation to sentences, we need to find the reduced parse appropriate for a particular sentence in the set obtained for it by the procedure described above. We use the EM selection algorithm proposed by Dębowski (2007).

The EM selection algorithm is an unsupervised statistic learning algorithm. Hence, its results depend on the size of a set of sentences it is applied on (for each verb separately). Thus, we decided to apply it to the whole set of sentences together. The sentences occurring in various sets were considered only once.

The application of the EM selection algorithm limits the number of sentences further. In the preparatory step it splits sentences into clauses (assigning corresponding reduced parses to them), increasing the number of sentences. However, it ignores all clauses with the number of reduced parses greater than 50. Furthermore, it ignores the verbs that have occurred only once. The impact of the algorithm on the statistics of sentences under consideration is presented in Table 6. We show the characteristics of source data (rows *source*), after preparatory steps (rows *prepare*) and final results of the EM algorithm (rows *final*).

Some reduced parses contain no arguments (cf. example (7)). Such reduced parses are deleted during EM algorithm preparatory step as well, unless they are the only reduced parse of a clause, which is a rather rare case in SEMKIPI. In

	all phrases						noun and prepositional phrases					
	d-Świgra			p-Świgra			d-Świgra			p-Świgra		
	mean	dev	med	mean	dev	med	mean	dev	med	mean	dev	med
SET 0	1.58	0.67	2	3.20	0.88	3	1.30	0.68	1	2.26	0.86	2
SET 1	1.63	0.68	2	3.15	0.85	3	1.34	0.71	1	2.32	0.90	2
SET 2	1.58	0.68	2	3.13	0.87	3	1.27	0.67	1	2.17	0.84	2
SET 3	1.58	0.67	2	3.22	0.88	3	1.30	0.68	1	2.30	0.88	2
SET 4	1.61	0.68	2	3.16	0.87	3	1.30	0.68	1	2.26	0.89	2
SET 5	1.54	0.67	2	3.17	0.86	3	1.26	0.65	1	2.39	0.96	2
SET 6	1.59	0.68	2	3.17	0.86	3	1.29	0.68	1	2.30	0.88	2
SET 7	1.64	0.68	2	3.19	0.89	3	1.32	0.69	1	2.24	0.88	2
SET 8	1.60	0.68	2	3.15	0.88	3	1.33	0.70	1	2.30	0.90	2
SET 9	1.59	0.68	2	3.19	0.88	3	1.32	0.70	1	2.34	0.92	2
TOTAL	1.59	0.68	2	3.17	0.87	3	1.30	0.69	1	2.29	0.89	2

TABLE 5: Statistics for the number of arguments of reduced parses

Table 6 we have added supplementary rows for source data, in which such "empty" reduced parses were not considered.

		sent-	reduced parses				phrases			noun & prep. phrs.		
		ences	nr.	mean	dev	med	mean	dev	med	mean	dev	med
d	source	19871	377811	19.0	19.6	7	1.74	0.66	2	1.46	0.72	1
i		——	359204	18.1	—	—	1.83	0.58	2	1.54	0.65	2
c	prepare	26330	158533	6.0	4.9	3	1.59	0.59	2	1.17	0.54	1
t	final	15685	53135	3.4	2.6	2	1.46	0.50	1	1.14	0.44	1
p	source	21378	4444720	207	271	34	3.26	0.90	3	2.27	0.88	2
e		——	4423263	207	—	—	3.27	0.89	3	2.28	0.87	2
r	prepare	21237	230197	10.8	8.6	6	2.08	0.70	2	1.18	0.65	1
m	final	12435	25630	2.1	1.2	1	1.32	0.44	1	0.95	0.44	1

TABLE 6: The statistics for the EM selection algorithm results

For *d-Świgra* parsed sentences, the preparatory step causes an increase in the number of sentences. This means that the division of sentences into clauses overbalanced the rejection of clauses with a large number of reduced parses. The fact that as many as 93% of sentences has at most 50 reduced parses (cf. Fig. 1) justifies this result. However, as many as $26330 - 15685 = 10645$ clauses were connected with verbs which occurred only once and were deleted during the main EM algorithm step. On the other hand, for *p-Świgra* parsed sentences, the preparatory step causes a small decrease of the number of sentences. The fact that only 58% of sentences has at most 50 reduced parses (cf. Fig. 1) balances the effect of division of sentences into clauses. The number of clauses connected with single-occurrence verbs rejected during the main EM algorithm step is smaller

(21237 − 12435 = 8802) than for *d-Świgra*, since many of them were deleted during the preparatory step. For final results, the number of sentences processed by both approaches is 10604, after *d-Świgra* parsing only is 5081 and after *p-Świgra* parsing only is 1831. Thus, the EM selection procedure for some sentences parsed by both versions of *Świgra* succeeds only for *d-Świgra* parsed sentences. The results of selection for *d-Świgra* parsed sentences will be referred to as *d-results*, whereas the results of selection for *p-Świgra* parsed sentences will be referred to as *p-results*.

It is obvious that the influence of the EM process on the number of reduced parses is much stronger for *p-results* than for *d-results*. Observe, however, that even though after the preparatory step mean as well as median of the number of reduced parses and phrases for *p-results* is still larger than for *d-results*, after the whole process it is not.

The EM selection algorithm selects one valence frame per clause. However, there may exist more than one reduced parse corresponding to one valence frame (cf. 7). Such ambiguity concerns each argument separately, hence in average the number of reduced parses is proportional to the number of arguments (phrases).

Concluding, the EM selection algorithm selects shorter valences frames for *p-Świgra* than for *d-Świgra*. It was implemented to choose a shorter frame from two or more equally probable frames. This makes an impression that such heuristic choice was more often applied to *p-Świgra* results.

All the above information says nothing about effectiveness of the whole process of parsing and selecting valence frames and reduced parses, i.e., we still do not know how the usage of a valence dictionary influences the number of sentences that have a proper reduced parse assigned. Unfortunately, the corresponding comparison of results cannot be performed automatically. Thus, we have selected 300 sentences belonging to *d-results* and *p-results*, 100 sentences belonging only to *d-results* and 60 sentences belonging only to *p-results*, proportionally to the sizes of the sets of parsed sentences. The corresponding sets of source reduced parses will be referred to as *d-common-source, p-common-source, d-only-source* and *p-only-source*, respectively. The corresponding sets of results of the EM selection algorithm will be referred to as *d-common, p-common, d-only* and *p-only*, respectively. These sets were evaluated manually.

Finally, to maximise the number of parsed sentences, using *p-Świgra* for sentences rejected by *d-Świgra* might be a good solution.

data	nr.	corr	acc	prec	rec	F
d-common	300	77.6	76.8	69.3	79.2	73.9
d-common (p)	300	77.6	90.5	69.3	79.2	73.9
p-common	300	74.2	90.3	64.6	73.1	68.6
d-only	100	70.5	73.5	50.2	69.1	58.1
p-only	60	58.3	90.8	53.5	85.9	65.9

Table 7: Evaluation of selected sets of sentences

Typically, measures elaborated for evaluation of algorithms are used to compare

the results of various algorithms run on the same data. This time we want to compare one and the same algorithm run on different data. This is especially important when we compare *d-common* and *p-common*. For every sentence, the set of reduced parses in *d-common-source* is a subset of the set of reduced parses in *p-common*. As a consequence, assuming the same choices, *p-common* would show the better *accuracy* than *d-common*, and the difference would enlarge with the growth of the number of errors. Therefore, we decided to present all most popular evaluation measures used in literature, namely *correctness, accuracy, precision, recall* and *F-measure*. They are presented in Table 7. In the case of *d-common* we did some trick concerning *accuracy* counting. Beside standard counting of this measure (cf. row *d-common* of the table), we counted it w.r.t. the number of reduced parses from *d-common-source* (cf. row *d-common (p)*). This means that we treat them as two methods run on the same data.

First observation is that the *d-common* and *p-common* sets contain the same results for 274 of 300 sentences (91%). This means that the algorithm behave very similarly in spite of the way the sentences were parsed. More precise measures for these sets show similar agreement, but all of them are a bit better for the *d-common* set (as for the *accuracy*, we should rather look at row *d-common (p)*).

The evaluation of the two other sets is evidently worse. However, remember that sentences in *d-only* are so complicated that *p-Świgra* did not manage to parse them or produced too many parses. Thus, the task of the EM selection algorithm was harder for them. On the other hand, the *p-only* set contains more sentences for which *Świgra* did not produce any valid reduced parse. This is indicated by low *correctness* value. However, this means that the number of true negatives is large, which increases *accuracy*, and the number of false negatives is small, which increases recall.

7 Conclusions

In this paper we have analysed the influence of using a valence dictionary in the process of parsing Polish sentences by means of *Świgra* parser. Valence information simplifies parsing process and hence leads to a decrease of resultant number of parse trees and reduced parse per sentence. Its disadvantage is that some correct sentences are not parsed (or no correct parse trees of a sentence and hence actual reduced parses of its clauses are obtained) because of the valence dictionary shortcomings. On the other hand, *d-Świgra* is able to parse more complicated sentences in a preset time limit.

The number of sentences accepted by *p-Świgra* is 2.8 percentage points greater than accepted by *d-Świgra*, but *p-Świgra* tends to choose longer reduced parses than *d-Świgra*. This means that it often treats adjuncts as complements.

All these results are stable. This property was tested on 10 sets of sentences randomly selected from the corpus.

The number of reduced parses is an important parameter for the EM selection algorithm, which is the next step of processing sentences in our project. As a result, we obtain a larger subcorpus of sentences with selected reduced parses for *d-Świgra* than for *p-Świgra*. Thus, using valence dictionary improves the process of

semantic annotation of sentences even from the perspective of the first two steps of this process. However, we should remember that the quality of the valence dictionary itself is an important factor.

Finally, a good solution to maximise the number of parsed sentences is to use *p-Świgra* to parse sentences rejected by *d-Świgra*.

References

Łukasz Dębowski (2007), Valence extraction using the EM selection and co-occurrence matrices, arXiv.

Łukasz Dębowski and Marcin Woliński (2007), Argument co-occurrence matrix as a description of verb valence, in Zygmunt Vetulani, editor, *Proceedings of the 3rd Language & Technology Conference*, pp. 260–264, Poznań, Poland.

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

Elżbieta Hajnicz (2009), Semantic annotation of verb arguments in shallow parsed Polish sentences by means of EM selection algorithm, in Małgorzata Marciniak and Agnieszka Mykowiecka, editors, *Aspects of Natural Language Processing*, volume 5070 of *LNCS*, Springer-Verlag.

Adam Przepiórkowski (2004), *The IPI PAN corpus. Preliminary version*, Institute of Computer Science, Polish Academy of Sciences, Warsaw.

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

Marcin Woliński (2005), An efficient implementation of a large grammar of Polish, *Archives of Control Sciences*, 15 (LI)(3):251—258.

Marcin Woliński (2006), Morfeusz — a Practical Tool for the Morphological Analysis of Polish, in Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors, *Proceedings of the Intelligent Information Systems New Trends in Intelligent Information Processing and Web Mining IIS:IIPWM'06*, Advances in Soft Computing, pp. 503–512, Springer-Verlag, Ustroń, Poland.

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

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