Towards the Automatic Acquisition of a Valence Dictionary for Polish

Adam Przepiórkowski[1,2]

[1] Institute of Computer Science, Polish Academy of Sciences,
ul. Ordona 21, 01-237 Warszawa, Poland,
adamp@ipipan.waw.pl,
http://nlp.ipipan.waw.pl/~adamp/
[2] Institute of Informatics, University of Warsaw,
ul. Banacha 2, 02-097 Warszawa, Poland

Abstract. This article presents the evaluation of a valence dictionary for Polish produced with the help of shallow parsing techniques and compares those results to earlier results involving deep parsing. We show that the valence dictionary obtained with the use of shallow parsing attains higher quality when it is measured on the basis of a corpus of valence frames, while the dictionary produced with the help of deep parsing seems superior when the results are compared to existing valence dictionaries.

Key words: valence acquisition, arguments of verbs, evaluation of valence dictionaries, partial parsing, IPI PAN Corpus of Polish

1 Aim and Scope

The valence of a given lexeme is, in general terms, its combinatorial potential, i.e., its ability to combine with other constituents of the utterance. In practice the term *valence* (or *valency*) usually refers to verbal lexemes, and it denotes the number and morphosyntactic makeup of the arguments of the verb. Hence, a valence dictionary will contain the information that the verb CHRAPAĆ, 'snore', combines only with the nominative subject, while the verb CHOWAĆ 'hide' also takes an accusative complement (*chować coś*, 'to hide something') and, optionally, a prepositional group with the preposition PRZED governing the instrumental case (*chować coś przed kimś*, 'to hide something from somebody'). Such dictionaries have various theoretical linguistic, psycholinguistic and educational uses, and they are a valuable resource in deep parsing and generation.

While there exist Polish dictionaries containing valence information,[3] including [15], [27], [2], and [14], the automatic acquisition of such information from corpora has many advantages when compared to the process of manual dictionary compilation. First, the automatic method is much faster and cheaper.

[3] A short comparison of a few of them in the context of Natural Language Processing may be found in [17], with some desiderata concerning such dictionaries put forward in [19].

M. Marciniak and A. Mykowiecka (Eds.): Bolc Festschrift, LNCS 5070, pp. 191–210, 2009.

Second, as it is based on naturally occurring texts, it is immune to prescriptive influences and to conflicting intuitions of a team of lexicographers. Hence, this method may be considered more objective. Third, the automatic procedure does not only extract valence frames, but also assigns them relative frequencies. For example, we may learn how often the verb DZIWIĆ 'make (one) wonder, surprise' combines with a nominal subject (*Dziwiło mnie takie postępowanie*, 'Such behaviour made me wonder'), and how often it co-occurs with a sentential subject (*Dziwiło mnie, że tak postąpił*, 'That he behaved like that made me wonder'). Such quantitative information is indispensable in probabilistic parsers (cf., e.g., [4], as well as [1]), which assign probabilities to particular parses, and it is also relevant in psycholinguistic research [13]. Fourth, as has been repeatedly noted (e.g., [28], [24], [12,11] and [9]), valence changes not only with time, but also with genre and topic. Once developed, automatic valence acquisition algorithms may be applied to various sets of texts in order to quickly and cheaply construct diachronic or thematic valence dictionaries. Fifth, automatic valence extraction may be used not only for the development of a new valence dictionary, but also for the verification and extension of an existing manually created dictionary (cf., e.g., [25]). [26] shows, for an automatically acquired valence dictionary of German, that the quality of such dictionaries may rival the quality of traditional dictionaries.

This paper focuses on the most basic type of valence information, which is found in all valence dictionaries, i.e., on morphosyntactic information regarding the grammatical class (part of speech) of the argument (e.g., CHOWAĆ, 'hide', combines with a nominal complement and not with a verbal complement), its grammatical case (e.g., CHOWAĆ combines with the accusative, not with the instrumental), etc. Some valence dictionaries, e.g., [15], also contain certain semantic information. Although the acquisition of such semantic valence information is beyond the scope of the current article, some work towards this end is currently being carried out within another ICS PAS project.[4]

2 Algorithm

As in virtually all previous work on valence acquisition (cf. [22], § 10.2, for an overview), the experiments described below proceed in two stages. First, at the linguistic stage, syntactic groups are identified which may be arguments of verbs. The result of this stage is a set of observations, where each observation consists of a verb and its observed potential arguments within a given sentence. Obviously, such observations will be noisy, with errors due to the inadequacies of morphosyntactic and syntactic processing. Second, the set of linguistic observations is subjected to statistical inference rules whose task is to decide which observations are reliable. Only thus filtered observations are considered valid valence frames.

[4] A Ministry of Science and Higher Education research grant (number NN516016533) "Automatyczne wykrywanie zależności semantycznych w strukturze argumentowej czasowników w dużych korpusach tekstów anotowanych syntaktycznie".

The main general steps of the algorithm, described in more detail below, are:

1. pre-process the empirical material, i.e., an appropriate subcorpus of the IPI PAN Corpus of Polish;
2. shallow process all sentences within that subcorpus and select fully parsed sentences for further statistical processing;
3. apply statistical filtering techniques, namely, the techniques proposed in [5].

2.1 Empirical Material

The main empirical material for the work reported here is the 2nd edition of the IPI PAN Corpus (`http://korpus.pl/`; [18]) containing about 255 million segments (over 200 million traditionally understood orthographic words, i.e., words delimited by white spaces and punctuation). Only sentences of less than 16 segments were selected from this corpus for further processing.[5] This restriction was imposed in order to reduce the time needed to process the massive amount of data, and similar restrictions are imposed in earlier work on the extraction of Polish valence, reported in [5]. Only candidates for true sentences, i.e., those containing verbal segments, were included in the processing chain.

As a result of this selection procedure, an IPI PAN subcorpus containing 25 647 017 segments (2 724 353 sentences) was created.

2.2 Shallow Parsing

The corpus obtained as described in the previous section was shallow parsed with the Spejd (`http://nlp.ipipan.waw.pl/Spejd/`; [23]) grammar presented in ch. 8 of [22], i.e., maximal nominal, prepositional, adjectival, verbal and other groups were automatically identified. Because of the partial nature of the grammar and the parser, not all sentences were fully parsed; after syntactic processing, some sentences contained sequences of lexical segments not assigned to any syntactic constituents. Only fully parsed sentences underwent the statistical processing. There were 1 137 014 (41.74%) such sentences and they contained 8 516 676 segments. One such sentence, *Kto się wstrzymał od głosu?* (‘Who abstained?’, literally: ‘Who self abstained from voice?’), is presented below:

```
<chunk type="s">
<group id="a106ac9" rule="(1) NG between verbs/groups/aby/etc."
       synh="a106abf" semh="a106abf" type="NG">
<tok id="a106abf">
<orth>Kto</orth>
<lex disamb="1">
 <base>kto</base><ctag>subst:sg:nom:m1</ctag>
</lex>
```

[5] 16 is the average length of a sentence in the 30-million “varied” (roughly balanced) subcorpus of the IPI PAN Corpus, `2.sample.30` (cf. `http://korpus.pl/index.php?page=download`).

```
</tok>
</group>
<group id="a106ac7" rule="sie" synh="a106ac0" semh="a106ac0"
       type="sie">
<tok id="a106ac0">
<orth>się</orth>
<lex disamb="1"><base>się</base><ctag>qub</ctag></lex>
</tok>
</group>
<syntok id="a106ac6" rule="czasownik niezanegowany 2a">
<orth>wstrzymał</orth>
<lex disamb="1">
 <base>wstrzymać</base><ctag>praet:sg:m1:perf:aff</ctag>
</lex>
<lex>
 <base>wstrzymać</base><ctag>praet:sg:m2:perf:aff</ctag>
</lex>
<lex>
 <base>wstrzymać</base><ctag>praet:sg:m3:perf:aff</ctag>
</lex>
<tok id="a106ac1">
<orth>wstrzymał</orth>
<lex disamb="1">
 <base>wstrzymać</base><ctag>praet:sg:m1:perf</ctag>
</lex>
<lex><base>wstrzymać</base><ctag>praet:sg:m2:perf</ctag></lex>
<lex><base>wstrzymać</base><ctag>praet:sg:m3:perf</ctag></lex>
</tok>
</syntok>
<group id="a106ac8"
       rule="(1) Dobre PrepNG na koncu zdania lub nawiasu"
       synh="a106ac2" semh="a106ac3" type="PrepNG">
<tok id="a106ac2">
<orth>od</orth>
<lex disamb="1"><base>od</base><ctag>prep:gen:nwok</ctag></lex>
</tok>
<tok id="a106ac3">
<orth>głosu</orth>
<lex disamb="1">
 <base>głos</base><ctag>subst:sg:gen:m3</ctag>
</lex>
</tok>
</group>
<ns/>
<tok id="a106ac5">
```

```
<orth>?</orth>
<lex disamb="1"><base>?</base><ctag>interp</ctag></lex>
</tok>
</chunk>
```

The syntactic representation exemplified above was subsequently translated to the format accepted by the statistical module, as proposed in [20] (and slightly modified in [5]). In the case of the above sentence, the result of this conversion is as follows:

```
% 'Kto się wstrzymał do głosu ?'
wstrzymać :np:nom: :prepnp:do:gen: :sie:
```

In the general case, the translation from the output format of Spejd to the input format of the statistical stage consists of the following steps:

1. each immediate constituent of a sentence, i.e., each XML child of the `<chunk type="s">` element, is assigned to one of the following three classes: grupa (i.e., a syntactic group), czasownik (i.e., a verb), inny token (neither, i.e., a token which is not a verb and does not belong to a recognised syntactic group); in particular, each token containing a verbal interpretation is assigned to the verbal class czasownik;
2. if, as a result, the number of elements in the verbal czasownik class is different than 1, the processing of the current sentence is aborted, and the algorithm moves to the next sentence;
3. since the only sentences that entered this stage of processing consisted of groups, verbs and punctuation marks, the class inny token must — after the previous steps — contain only punctuation marks and, as such, it is ignored in further processing;
4. the orthographic makeup of the sentence is retrieved for the purpose of a comment in the resulting file (starting with a `%`; cf. the example above);
5. the base form of the single verb in the sentence is retrieved; if this segment has a number of verbal interpretations with different base forms, the first of them is arbitrarily assumed to be the correct one;
6. each syntactic group belonging to the grupa class is translated into a list of morphosyntactic interpretations of the syntactic head of the group, e.g., `:np:nom:`, `:np:acc:`, `:prepnp:do:gen:`, `:infp:imperf:`; as a head may contain a number of morphosyntactic interpretations, the result is a list rather than a single such representation;
7. the Cartesian product of the lists (treated as sets) of representations of all elements of the grupa class is taken as the set of potential observations adduced by the currently processed sentence;
8. the potential observations are sorted and printed out.

Despite the fact that the shallow Spejd grammar used in the experiments reported in this paper contains some morphosyntactic disambiguation rules, not all segments are fully disambiguated, so one sentence may be the basis of a number of different potential observations, calculated in steps 6–7 of the above algorithm. For example, 4 potential observations were obtained for the sentence *Składam te podziękowania na ręce szefowej komisji pani senator Genowefy Grabowskiej* ('I thank the head of the commission, senator Genowefa Grabowska', literally: 'I-put these thanks onto hands boss.`gen` commission.`gen` Mrs. senator Genowefa Grabowska'):[6]

```
% 'Składam te podziękowania na ręce szefowej komisji
%  pani senator Genowefy Grabowskiej .'
składać :np:acc: :prepnp:na:acc:
składać :np:acc: :prepnp:na:loc:
składać :np:nom: :prepnp:na:acc:
składać :np:nom: :prepnp:na:loc:
```

The group *te podziękowania* 'these thanks' is not fully disambiguated and it retains both the accusative and the nominative case interpretations, and similarly the prepositional group, *na ręce...* 'onto hands', is not disambiguated as to the accusative or locative case value, which results in 4 potential observations. Note that, unlike in the case of the deep parser Świgra (`http://nlp.ipipan.waw.pl/~wolinski/swigra/`; [30,31]) utilised in [5], the observations may only differ in the values of morphosyntactic categories, not in the number or extent of syntactic groups; following the general shallow parsing principles, Spejd outputs a unique parse of the sentence.

A more crucial difference between the current experiments and the approach proposed by Dębowski consists in our refraining from any further linguistic processing: all linguistic knowledge is contained in the grammar, and the resulting observations correspond directly to the groups found by the parser. This should be contrasted with the algorithm described in [5], where the results of the grammar are subject to some further linguistic processing, including the following steps:

- a nominal group is added to an observation in case an elided subject (so-called *pro*-drop) is detected;
- the nominal genitive group, if any, is removed from an observation in case of a negated sentence, as this genitive group may actually be a Genitive of Negation (cf. [16]) realisation of an otherwise accusative argument of the verb;
- nominal phrases suspected of having the grammatical function of (temporal) adjuncts are removed from observations.

It is not clear to what extent such further transformations influenced the final results of [5], but they probably played a role in producing results more comparable to valence frames found in existing valence dictionaries. Such *a posteriori* modifications of observations must also lead to less accurate data concerning the actual text frequencies of particular realisations of valence frames.

[6] The orthographic form of the sentence, given here as a comment, was broken for typographical reasons.

2.3 Statistical Processing

The pre-processing step of the statistical stage is the selection of a single observation in case of sentences with multiple potential observations. A simple EM-type (Expectation Maximisation) algorithm described in [5] is used to this end. In the case of the example sentence given above, the observation actually selected for *Składam te podziękowania na ręce szefowej komisji pani senator Genowefy Grabowskiej* correctly assumes the accusative case of the nominal group, but incorrectly identifies the case within the prepositional group as locative:

```
% 'Składam te podziękowania na ręce szefowej komisji
%  pani senator Genowefy Grabowskiej .'
składać :np:acc: :prepnp:na:loc:
```

Observations collected and further selected this way are the first version of the resulting valence dictionary, the so-called proto-dictionary [7]. For example, the lexical entry for the verb WYPŁYWAĆ, 'flow out, emerge, follow' in the proto-dictionary created within the current shallow parsing experiments is given below:[7]

```
'wypływać' => {
  'np(nom),z+np(gen)' => 9,
  'adv,np(nom),z+np(gen)' => 2,
  '' => 1,
  'adj(nom),adv,dla+np(gen),np(acc)' => 1,
  'adj(nom),np(acc),z+np(gen)' => 1,
  'adj(nom),z+np(gen)' => 1,
  'adv' => 1,
  'dla+np(gen),np(nom),z+np(gen)' => 1,
  'do+np(gen),np(nom),z+np(gen)' => 1,
  'np(acc),o+np(loc),z+np(gen)' => 1,
  'np(acc),od+np(gen),z+np(gen)' => 1,
  'np(dat),np(nom)' => 1,
  'np(nom),np(voc),z+np(gen)' => 1,
  'np(nom),o+np(loc)' => 1
}
```

According to this entry, forms of the verb WYPŁYWAĆ were observed 9 times with a nominative nominal group and a prepositional group headed by the genitive-taking preposition Z, twice with an additional adverbial group, once without any accompanying groups, etc.

The two main steps of the statistical stage make use of an approximate representation of valence proposed in [6], where a valence frame is described as a *set* of possible arguments of the verb (the set of all arguments in all possible frames of that verb) and an additional table specifying whether any two possible

[7] The format of such lexical entries is actually the representation of hash tables in the Perl programming language.

arguments always co-occur, never co-occur, unidirectionally imply one another, or seem independent of each other.

First, all possible arguments of a given verb are collected. A subset of that argument set is identified as those arguments which occur in all possible frames of that verb. An argument type a is considered a possible argument of a verb v in case the inequality in (1) holds; $c(v)$ denotes here the number of occurrences of the verb v, $c(v, a)$ — the number of observed co-occurrences of v with the argument a. The argument is additionally considered a necessary argument of v in case (2) holds.

(1) $c(v, a) \geq p_a c(v) + 1$
(2) $c(v) - c(v, a) < p_{\neg a} c(v) + 1$

The parameters p_a and $p_{\neg a}$ occurring in the inequalities above are trained — separately for each argument type — on the basis of the manually created dictionary [27], as well as on the basis of lexical entries of around 200 verbs in [15] and [2]. The exact parameter estimation procedure for p_a and $p_{\neg a}$ is described in [5].

This first step ends with projecting information of estimated possible and necessary arguments into actually observed frames: in each observed frame only those arguments are retained which are "possible" in the sense above and, moreover, in case the observation does not contain the "necessary" argument, it is artificially added to the frame.[8] As a result of this step, the lexical entry of WYPŁYWAĆ is reduced as follows:

```
'wypływać' => {
  'np(nom),z+np(gen)' => 15,
  'np(nom)' => 4,
  'np(acc),np(nom),z+np(gen)' => 3,
  'np(acc),np(nom)' => 1
}
```

It follows from the comparison of this lexical entry with the corresponding lexical entry in the proto-dictionary that `adv`, `np(voc)`, `np(dat)` and various prepositional argument types were rejected as possible arguments of the verb, so the set of possible arguments is reduced to {`np(nom)`, `np(acc)`, `z+np(gen)`}. Further, the nominative nominal group was classified as a necessary argument. Hence, the four "observations" `'np(nom)'` in the lexical entry above actually correspond to the original observations: `'np(dat),np(nom)'`, `'np(nom),o+np(loc)'`, `'adv'` and the empty observation `''`.

In the second step, full frames obtained in the first step are evaluated. Again, on the basis of existing valence dictionaries, possible relationships between arguments are estimated: do the two given arguments usually co-occur within frames of various verbs, do they have a complementary distribution, does one imply the other, or are they independent. For any two argument types, such a

[8] This description is based on the observation of the algorithm at work and it differs a little from the description in [5].

relationship is calculated on the basis of the whole dictionary, independent of particular verbs. For a given verb, this default relation between two arguments is assumed, unless there are strong reasons to override it. For example, the 15 “observations” of the nominative nominal group `np(nom)` co-occurring with the `z+np(gen)` prepositional group in the vicinity of WYPŁYWAĆ were not sufficient to retain that frame of that verb (note that the frame is correct here, but it is generally rather rare in the corpus), so the final lexical entry for WYPŁYWAĆ looks as follows:

```
'wypływać' => {
  'np(nom)' => 4,
  'np(acc),np(nom),z+np(gen)' => 3,
  'np(acc),np(nom)' => 1
}
```

3 Results

Three dictionaries were created as a result of the algorithm described in the previous section: the proto-dictionary, which contains the actual observations (perhaps selected from alternative potential observations with the help of a simple EM algorithm), the dictionary resulting from the first step of statistical processing (the intermediate dictionary), and the final dictionary created in the second step of statistical processing. Table 1 gives the sizes of these dictionaries.

Table 1. Sizes of automatically obtained valence dictionaries

dictionary	**entries**	**frames**	
		tokens	types
proto	6,845	1,084,286	20,894
intermediate	6,845	1,084,286	517
final	4,166	863,731	141

The proto-dictionary obtained as described in the previous sections contains 6 845 entries. As a result of the second step of statistical processing, this number is reduced to 4 166 entries with 207.33 observations per entry on average. The full dictionary would consume around 430 pages, so — given the space limits — it must suffice to present some of its characteristics here.[9]

The final number of different “observed” valence frames is 141. This number is substantially reduced with respect to the number of 20 894 different realisations of frames actually observed, and also with respect to the 517 types of “observations” remaining after the first step of statistical processing. The most

[9] Appendix A contains a fragment of the dictionary resulting from the simplification of the statistical stage, as described in § 5.

frequent frame was the intransitive frame (only a nominative nominal group; 232 034 occurrences) and the empty frame (129 720), while the actually very frequent transitive frame (nominative and accusative nominal groups) is the 4th most frequent frame in the resulting final dictionary (84 611 occurrences). Out of the three frames with single occurrences: `'do+np(gen),np(gen),np(nom),sie'` (for the verb UŻYWAĆ 'use'), `'np(dat),np(nom),o+np(loc),sie'` (MARZYĆ 'dream') i `'np(nom),o+np(loc),z+np(inst)'` (POROZMAWIAĆ 'talk'), the first two are erroneous, and the last one seems to be correct.

4 Evaluation

Two evaluation procedures were applied to the dictionaries obtained as described above: the dictionary-based evaluation (at the level of frame types) and the corpus evaluation (at the level of frame occurrences, or tokens).

4.1 Dictionary-Based Evaluation

Dictionary-based evaluation consists in finding the ratio of automatically extracted valence frames also present in manually constructed dictionaries (precision), and the ratio of frames in such previously built dictionaries also present in the automatic results (recall). In the current experiments, these values were estimated on the basis of a sample of 202 verbs randomly selected from [27]. For all these 202 verbs, their entries were also extracted from two other manually constructed dictionaries [15,2] and converted to the "least common denominator" format.[10] Since the dictionaries differed a little in the scope and character of the valence information, the conversion process was to some extent interpretative.

Precision (P), recall (R) and their harmonic mean, called F-measure (F), were computed for the final dictionary obtained in the current experiments, as well as for both dictionaries reflecting earlier stages of processing: for the proto-dictionary and for the intermediate dictionary. In each case automatically obtained valence frames were compared to various gold standards, that is, to each of the three manually constructed valence dictionaries, marked below as "Bań." [2], "Pol." [15] and "Świ." [27], and to two dictionaries compiled from these three manually constructed dictionaries by taking their set-theoretical sum ("SUM") and by majority voting ("MV"; i.e., a frame is present in the MV dictionary, if it is present in at least two manually constructed dictionaries). In each comparison, only the frames of those verbs were considered which were present both in the automatically obtained dictionary and in the gold standard. The results of these comparisons are contained in Tables 2–4.

The comparison of Tables 2 and 3 shows the great importance of the first step of statistical processing. Although it resulted in a significant drop of recall (from 40.40% to 29.80%, for the MV dictionary), that decrease in recall was

[10] These data were prepared by Witold Kieraś and Łukasz Dębowski, whose scripts were used for calculating precision and recall given below.

Table 2. Dictionary-based evaluation of the proto-dictionary

	Bań.	Pol.	Świ.	SUM	MV
P	3.97	3.04	3.05	5.11	**3.15**
R	37.33	31.03	34.83	28.60	**40.40**
F	7.17	5.54	5.62	8.68	**5.85**

Table 3. Dictionary-based evaluation of the intermediate dictionary

	Bań.	Pol.	Świ.	SUM	MV
P	50.69	41.70	39.95	57.80	**44.94**
R	24.68	22.00	23.56	16.72	**29.80**
F	33.20	28.81	29.64	25.94	**35.84**

Table 4. Dictionary-based evaluation of the final dictionary

	Bań.	Pol.	Świ.	SUM	MV
P	63.81	52.23	51.34	70.41	**57.58**
R	22.04	19.51	21.49	14.42	**27.07**
F	32.77	28.41	30.30	23.94	**36.83**

more than compensated by the dramatic increase in precision (from 3.15% to 44.94%), which resulted in the clear increase in the harmonic mean of these measures (from 5.85 to 35.84). The second step of the statistical stage wasn't so significant: although the F-measure for the MV dictionary is higher in Table 4 than in Table 3, the difference is relatively minor (36.83 to 35.84) because the increase in precision was to a large extent annulled by the decrease in recall. It is interesting to note that, for two gold standards, Bań. and Pol., the second step of statistical processing was slightly detrimental, if the quality were evaluated with the F-measure.

Let us also note at the end of this section that, although the results are far from perfect, they constitute a clear improvement over the reasonable baseline consisting in the assignment of two frames to each verb: the intransitive frame `'np(nom)'` and the transitive frame `'np(acc),np(nom)'`.[11] A "dictionary" constructed this way would have relatively high precision (47.41% when measured against the MV dictionary), but very low recall (15.15%). Complete results of the evaluation of such a baseline dictionary are given in Table 5.

4.2 Corpus-Based Evaluation

Also corpus-based evaluation shows that, after shallow processing at the linguistic stage, it may be beneficial to stop statistical processing after the first step,

[11] Experiments were also performed for other baselines, including: only the intransitive frame, only the transitive frame, the empty frame, the infinitival frame, and various combinations of these frames. In each case the dictionary-based evaluation gave worse (in terms of F-measure) results than the results for the baseline given below.

Table 5. Dictionary-based evaluation of the baseline, i.e., a dictionary created artificially by assuming two frames for each verb: the transitive frame and the intransitive frame

	Bań.	Pol.	Świ.	SUM	MV
P	54.66	42.49	43.52	59.33	**47.41**
R	12.83	10.80	12.37	8.27	**15.15**
F	20.78	17.23	19.27	14.52	**22.96**

before frames are removed on the basis of the low co-occurrence of the arguments in manually constructed dictionaries.

The evaluation was performed for 12 verbs selected on the basis of their frequencies in the corpus resulting from the linguistic processing. These are 4 very frequent verbs (tens of thousands of occurrences): WSTRZYMAĆ 'stop', CHCIEĆ 'want', STWIERDZAĆ 'conclude', MIEĆ 'have', 4 verbs of medium frequency (around 4 000 occurrences): MUSIEĆ 'must', ZABRAĆ 'take (away)', PRZYPOMINAĆ 'remind, remember', and ZGŁOSIĆ 'report', and 4 relatively rare verbs (around 400 occurrences): STAWIAĆ 'put', USŁYSZEĆ 'hear', ZAUWAŻYĆ 'notice' and USTALIĆ 'establish'.

For each verb, 120 sentences containing that verb were randomly selected. These sentences were linguistically annotated[12] on the basis of brief guidelines of syntactic annotation [21] with the help of the Anotatornia annotation tool developed at ICS PAS [8].

Some of these sentences were lacking full morphosyntactic analysis, some were the result of erroneous segmentation of text into sentences. The remaining 985 sentences were annotated for maximal syntactic groups. Hence, the result of the annotation was a set of fully correct valence frame observations. Obviously, just as would be the case in fully correct shallow processing, these observations contained information of all observed dependents of the main verb: arguments and adjuncts alike. Also, these observations were not further processed linguistically in any way; in particular, no information about missing (elided) arguments was added.

The corpus prepared this way was the basis for calculating token recall, i.e., the ratio of manually annotated frames also found by the algorithm. The result was 89% for the proto-dictionary, 32% for the intermediate dictionary and 22% for the final dictionary. Table 6 presents the results in more detail, while Table 7 presents analogous results for the approach based on deep linguistic processing, described in [5]. It is interesting to note that, although the proto-dictionary based on shallow processing with Spejd contains many more valence frames observed in manually annotated texts than the proto-dictionary based on the deep parser Świgra (89% compared to 39%), this difference reduces significantly for the intermediate dictionary (32% to 27%) and reverses for the final dictionary (22% to 27%). This effect is probably to some extent caused by the greater

[12] By linguists: Monika Czerepowicka, Hanna Maliszewska, Marta Nazarczuk-Błońska, Marta Piasecka and Izabela Will.

Table 6. The number of observations of valence frames for the 12 verbs for which the appropriate frame is also present in the valence dictionary automatically obtained with the use of the Spejd parser and the grammar presented in ch. 8 of [22]

verb	frames (tokens)			
		in dictionary		
	in texts	proto	intermediate	final
USTALIĆ	73	54	6	11
ZABRAĆ	103	100	1	1
STAWIAĆ	78	34	8	6
CHCIEĆ	91	89	19	19
ZAUWAŻYĆ	65	48	6	11
WSTRZYMAĆ	88	88	88	88
MIEĆ	86	84	28	28
MUSIEĆ	84	83	34	34
PRZYPOMINAĆ	108	93	0	0
STWIERDZAĆ	119	119	114	0
ZGŁOSIĆ	73	70	15	15
USŁYSZEĆ	17	13	0	0
total	985	875	319	213
percent	100	88.83	32.39	21.62

Table 7. The number of observations of valence frames for the 12 verbs for which the appropriate frame is also present in the valence dictionary automatically obtained with the use of the Świgra deep parser

verb	frames (tokens)			
		in dictionary		
	in texts	proto	intermediate	final
USTALIĆ	73	23	8	10
ZABRAĆ	103	93	53	53
STAWIAĆ	78	22	8	6
CHCIEĆ	91	23	19	19
ZAUWAŻYĆ	65	18	7	11
WSTRZYMAĆ	88	88	88	88
MIEĆ	86	43	28	28
MUSIEĆ	84	37	34	34
PRZYPOMINAĆ	108	8	7	4
STWIERDZAĆ	119	0	0	0
ZGŁOSIĆ	73	23	18	10
USŁYSZEĆ	17	2	0	0
total	985	380	270	263
percent	100	38.58	27.41	26.70

dispersion of data in the current approach (many more different valence frame types are found in shallow processing), but it may also be a result of the close fit of the statistical approach proposed by Dębowski and the deep parsing with Świgra, on the basis of which that approach was developed and fine-tuned [6,5].

As shown in the next section, the rather disappointing results given above improve when the second step of statistical processing is simplified. Nevertheless, already these modest results are well above the baseline described above: in the case where every verb is assigned the transitive frame and the intransitive frame, the resulting valence dictionary would cover only 101 (10.25%) corpus observations.

5 Simplification of Statistical Processing

The previous three sections describe some experiments in valence extraction, where linguistic processing is performed with the Spejd implementation of the shallow grammar presented in ch. 8 of [22], while the statistical processing follows the ideas described in [5]. In the preceding section we noted that the second step of the statistical stage, where frames with uncommon combinations of arguments are rejected, is a mixed blessing at best: it improves the results of the dictionary-based evaluation only slightly (and in fact has a detrimental effect, if Polański's or Bańko's dictionaries are taken as gold standards), and it causes a clear drop in the quality measured via corpus-based evaluation.

On the other hand, as noted in various earlier works on valence acquisition for other languages (e.g., [3,10,11,26,5]), simpler methods of rejecting rare observations often give results comparable to more complicated statistical techniques. Hence, it would be interesting to find out whether applying such simpler methods in the current linguistic setup also gives results comparable to or better than the techniques proposed in [5].

Table 8. Dictionary-based evaluation (for the MV dictionary) of valance information acquired by rejecting observations rare in the intermediate dictionary; for comparison, the table also recalls previous results for the intermediate and final dictionary

$c_{\min}(v)$	10	10	13	**dictionary**	
$p_{\min}(r, v)$	0	2	2	**intermediate**	**final**
P	45.49	53.08	53.01	44.94	57.58
R	32.45	30.94	31.45	29.80	27.07
F	37.88	39.09	39.48	35.84	36.83

To this end, further experiments based on shallow linguistic processing were conducted, where the second step of statistical processing was replaced with a simpler rejection of rare observations. Two parameters, or cutoff points, were used: the sheer number of occurrences of the verb, $c_{\min}(v)$, and the ratio of the number of co-occurrences of a given frame with a given verb to the numer

Table 9. Dictionary-based evaluation of valance information acquired with shallow linguistic processing and with cutoff points $c_{\min}(v) = 13$ and $p_{\min}(r, v) = 2$

	Bań.	Pol.	Świ.	SUM	MV
P	58.53	49.00	46.66	66.05	**53.01**
R	25.13	23.05	25.09	16.98	**31.45**
F	35.16	31.35	32.63	27.02	**39.48**

of all occurrences of that verb, $p_{\min}(r, v)$ (expressed as percent points). The requirement that valence frames be acquired only for verbs occurring at least $c_{\min}(v) = 10$ times in the parsed corpus improved F-measure (as computed for the MV dictionary) to 37.88, and further rejection of observations less frequent than $p_{\min}(r, v) = 2$ (i.e., 2%) increased the value to 39.09. In various experiments performed, the best F-measure, 39.48, was achieved for $c_{\min}(v) = 13$ and $p_{\min}(r, v) = 2$. The results are summarised and compared to earlier results in Table 8, while more complete results for the best cutoff points are given in Table 9.

Let us note that significant improvements as measured by dictionary-based evaluation were achieved with practically no decrease in the quality measured with corpus-based evaluation (cf. Table 10). The number of corpus observations corresponding to automatically identified frames is 318, i.e., almost the same as in the intermediate dictionary (319; cf. Table 6 on p. 203), and much higher than in the final dictionary (213).

Table 10. The number of observations of valence frames for the 12 verbs for which the appropriate frame is also present in the valence dictionary automatically obtained with the use of the Spejd parser and the grammar presented in ch. 8 of [22] (simplified statistical processing)

	frames (tokens)	
verb	**in texts**	**in dictionary**
USTALIĆ	73	6
ZABRAĆ	103	1
STAWIAĆ	78	7
CHCIEĆ	91	19
ZAUWAŻYĆ	65	6
WSTRZYMAĆ	88	88
MIEĆ	86	28
MUSIEĆ	84	34
PRZYPOMINAĆ	108	0
STWIERDZAĆ	119	114
ZGŁOSIĆ	73	15
USŁYSZEĆ	17	0
total	985	318
percent	100	32.28

Table 11. Dictionary-based evaluation of valance information acquired with deep linguistic processing [5]

	Bań.	Pol.	Świ.	SUM	MV
P	63.53	54.56	55.17	74.01	**59.88**
R	25.39	23.63	26.71	17.58	**32.59**
F	36.28	32.98	35.99	28.41	**42.21**

Table 12. Dictionary-based evaluation of valance information acquired with deep linguistic processing and with cutoff points $c_{\min}(v) = 17$ and $p_{\min}(r, v) = 2$

	Bań.	Pol.	Świ.	SUM	MV
P	59.26	49.66	49.66	69.82	**54.18**
R	27.98	25.86	29.87	20.00	**35.55**
F	38.01	34.01	37.30	31.09	**42.93**

It should also be noted that replacing the second statistical step with cutoff points in the original methodology — based on deep parsing with the Świgra parser — described in [5] also brings about certain, but less significant, improvements in the values of the F-measure. In this case the best cutoff points were $c_{\min}(v) = 17$ and, as above, $p_{\min}(r, v) = 2$. The results of dictionary-based evaluation for these cutoff values are given in Table 12, while the original results presented in [5] are cited in Table 11. When compared to the results in Table 7, the result of corpus-based evaluation is practically the same: there were 264 corpus observations corresponding to automatically acuired valence frames.

Taking into consideration both evaluation methodologies, the results based on shallow linguistic processing are comparable to Dębowski's ([5]) results based on deep processing; such a comparison is presented in Table 13. While the current experiments produce inferior results, when measured as the similarity to the majority voting dictionary, they are clearly superior when measured with reference to actually occurring frame realisations of the 12 verbs of varying frequencies.

6 Summary

The aim of this article was to present a practical application of the formalism and the grammar described in [22] to the task of automatic valence acquisition from morphosyntactically annotated corpora. The quality of the results of valence acquisition with shallow parsing and simplified statistical processing turns out to be comparable to the best results for Polish found in the literature, and much higher when measured against frames actually observed in texts. Also, the simplification of the statistical stage alone makes it possible to slightly improve

Table 13. A comparison of final results of three approaches: [5], the approach presented there with the second step of statistical processing replaced by simple cutoff points, and the approach presented in [22] and summarised in this article, also with simple cutoff points instead of the second step of statistiacal processing; P, R and F are precision, recall and their harmonic mean, as measured in dictionary-based evaluation, and C is the corpus-based token recall; the best results are in boldface

	Dębowski ([5])	Dębowski ([5]) $c_{\min}(v) = 17$ $p_{\min}(r, v) = 2$	Przepiórkowski ([22]) $c_{\min}(v) = 13$ $p_{\min}(r, v) = 2$
P	**59.88**	54.18	53.01
R	32.59	**35.55**	31.45
F	42.21	**42.93**	39.48
C	26.70	26.80	**32.28**

the results of the dictionary-based evaluation, when compared to the earlier best results described in [5].

There are many possible ways the approach presented above may be developed further and improved. The most obvious concern linguistic processing: both the morphological analyser and the shallow grammar could be extended in various ways. Also the empirical basis could be improved, not only by increasing the corpus size, but also by making better use of the current corpus: at the moment evidence provided by subordinate clauses and less than fully parsed sentences is lost in the process. The evaluation of the results obtained using different linguistic and statistical methods also suggests that the novel approach to the statistical stage proposed in [5], promising in combination with deep processing at the linguistic stage, may be less adequate when coupled with shallow linguistic processing. We hope to continue work both on the empirical basis and on linguistic and statistical methodologies of valence acquisition within subsequent projects carried out at ICS PAS.

Acknowledgements. This article contains some of the material of chapter 10 of the monograph [22], summarising the main results of a valence acquisition project carried out at the Institute of Computer Science of the Polish Academy of Sciences (ICS PAS) in 2005–2008.[13] The acknowledgements therein carry over to this paper, with additional thanks for comments to Małgorzata Marciniak.

A An Extract from the Valence Dictionary

This appendix contains an extract from the valence dictionary automatically acquired with the use of the shallow Spejd grammar presented in ch. 8 of [22], combined with the simplified statistical processing described in § 5.

[13] A Ministry of Science and Higher Education research grant (number 3T11C00328) “Automatyczna ekstrakcja wiedzy lingwistycznej z dużego korpusu języka polskiego”, http://nlp.ipipan.waw.pl/PPJP/.

```
'gadać' => {
 'np(nom)' => 58,
 'np(acc),np(nom)' => 8,
 'np(nom),PZ' => 5
}
'gasić' => {
 'np(nom)' => 12,
 'np(acc),np(nom)' => 5
}
'gasnąć' => {
 'np(nom)' => 10,
 'nad+np(inst),np(nom)' => 3
}
'generować' => {
 'np(nom)' => 12,
 'np(acc),np(nom)' => 6
}
'ginąć' => {
 'np(nom)' => 49
}
'gniewać' => {
 'np(nom),sie' => 22,
 'np(nom),sie,ZE' => 3
}
'godzić' => {
 'na+np(acc),np(nom),sie' => 42,
 'inf,np(nom),sie' => 28,
 'np(nom),w+np(acc)' => 26,
 'np(nom),sie' => 17,
 'np(acc),np(nom),w+np(acc)' => 4,
 'np(acc),np(nom)' => 4,
 'np(acc),np(nom),sie' => 4,
 'np(nom)' => 3
}
'gonić' => {
 'np(acc),np(nom)' => 17,
 'np(nom)' => 15
}
'gospodarować' => {
 'np(inst),np(nom)' => 7,
 'np(nom)' => 7,
 'np(acc),np(nom)' => 2,
 'np(acc),np(inst),np(nom)' => 1
}
'gotować' => {
 'np(acc),np(nom)' => 4,
 'np(nom)' => 3,
 'do+np(gen),np(nom),sie' => 3,
 'np(nom),sie' => 3,
 'do+np(gen),np(acc),np(nom),sie'
      => 1,
 'np(dat),np(nom),sie' => 1,
 'np(dat),np(nom)' => 1
}
'gościć' => {
 'np(nom)' => 24,
 'np(acc),np(nom)' => 14
}
'gratulować' => {
 'np(nom)' => 97,
 'np(dat),np(nom)' => 39
}
'grać' => {
 'np(nom)' => 267,
 'np(acc),np(nom)' => 21
}
'gromadzić' => {
 'np(nom)' => 10,
 'np(nom),sie' => 10,
 'np(acc),np(nom)' => 8,
 'np(acc),np(nom),sie' => 4,
 'na+np(acc),np(acc),np(nom)' => 3,
 'na+np(acc),np(nom)' => 1,
 'na+np(acc),np(nom),sie' => 1
}
'grozić' => {
 'np(dat),np(nom)' => 83,
 'np(inst),np(nom)' => 54,
 'np(nom)' => 38,
 'do+np(gen),np(dat),np(nom)' =>
      11,
 'do+np(gen),np(dat),np(nom),
   za+np(acc)' => 10,
 'do+np(gen),np(acc),np(dat),
   np(nom)' => 7,
 'np(dat),np(inst),np(nom)' => 6
}
'gubić' => {
 'np(nom),sie' => 12,
 'np(nom)' => 6,
 'np(acc),np(nom)' => 4,
 'np(acc),np(nom),sie' => 2
}
'gwarantować' => {
 'np(acc),np(nom)' => 55,
 'np(nom)' => 41,
 'np(nom),ZE' => 26,
 'np(acc),np(dat),np(nom)' => 8,
 'np(dat),np(nom)' => 6,
 'np(dat),np(nom),ZE' => 4
}
```

References

1. Arun, A., Keller, F.: Lexicalization in crosslinguistic probabilistic parsing: The case of French. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, MI, pp. 306–313 (2005)
2. Bańko, M. (ed.): Inny słownik języka polskiego. Wydawnictwo Naukowe PWN, Warsaw (2000)
3. Briscoe, T., Carroll, J.: Automatic extraction of subcategorization from corpora. In: Proceedings of the 5th Applied Natural Language Processing Conference, Washington, DC, pp. 356–363. ACL (1997)
4. Carroll, J., Minnen, G., Briscoe, T.: Can subcategorisation probabilities help a statistical parser? In: Proceedings of the Sixth Workshop on Very Large Corpora, Montreal, pp. 118–126 (1998)
5. Dębowski, Ł.: Valence extraction using the EM selection and co-occurrence matrices, 5 Dec. 2007. arXiv:0711.4475v2 [cs.CL] (2007)
6. Dębowski, Ł., Woliński, M.: Argument co-occurence matrix as a description of verb valence. In: Vetulani, Z. (ed.) Proceedings of the 3rd Language & Technology Conference, Poznań, Poland, pp. 260–264 (2007)
7. Dębowski, Ł., Woliński, M.: Nowe metody ekstrakcji walencji czasowników z tekstów w języku polskim. Referat wygłoszony na seminarium Zespołu Inżynierii Lingwistycznej IPI PAN, Warszawa, 22 października 2007 (2007)
8. Hajnicz, E., Murzynowski, G., Woliński, M.: ANOTATORNIA – lingwistyczna baza danych. In: Materiały V konferencji naukowej InfoBazy 2008, Systemy * Aplikacje * Usługi, Gdańsk, pp. 168–173. Centrum Informatyczne TASK, Politechnika Gdańska (2008)
9. Hare, M., McRae, K., Elman, J.: Sense and structure: Meaning as a determinant of verb subcategorization preferences. Journal of Memory and Language 48(2), 281–303 (2003)
10. Kawahara, D., Kaji, N., Kurohashi, S.: Japanese case structure analysis by unsupervised construction of a case frame dictionary. In: Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000), Saarbrücken, pp. 432–438 (2000)
11. Korhonen, A.: Subcategorization Acquisition. Ph. D. dissertation, University of Cambridge (2002)
12. Korhonen, A.L.: Using semantically motivated estimates to help subcategorization acquisition. In: Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, ACL (2000)
13. Lapata, M., Keller, F., Schulte im Walde, S.: Verb frame frequency as a predictor of verb bias. Journal of Psycholinguistic Research 30(4), 419–435 (2001)
14. Mędak, S.: Praktyczny słownik łączliwości składniowej czasowników polskich. Universitas, Cracow (2005)
15. Polański, K. (ed.): Słownik syntaktyczno-generatywny czasowników polskich. Zakład Narodowy im. Ossolińskich / Instytut Języka Polskiego PAN, Wrocław / Cracow (1992)
16. Przepiórkowski, A.: Long distance genitive of negation in Polish. Journal of Slavic Linguistics 8, 151–189 (2000)
17. Przepiórkowski, A.: On the computational usability of valence dictionaries for Polish. IPI PAN Research Report 971, Institute of Computer Science, Polish Academy of Sciences, Warsaw (2003)

18. Przepiórkowski, A.: The IPI PAN Corpus: Preliminary version. Institute of Computer Science. Polish Academy of Sciences, Warsaw (2004)
19. Przepiórkowski, A.: Towards the design of a syntactico-semantic lexicon for Polish. In: Kłopotek, M.A., Wierzchoń, S.T., Trojanowski, K. (eds.) Intelligent Information Processing and Web Mining. Advances in Soft Computing, pp. 237–246. Springer, Berlin (2004)
20. Przepiórkowski, A.: What to acquire from corpora in automatic valence acquisition. In: Koseska-Toszewa, V., Roszko, R. (eds.) Semantyka a konfrontacja językowa, vol. 3, pp. 25–41. Slawistyczny Ośrodek Wydawniczy PAN, Warsaw (2006)
21. Przepiórkowski, A.: Krótka instrukcja anotacji składniowej. Unpublished manuscript, Institute of Computer Science, Polish Academy of Sciences (2008)
22. Przepiórkowski, A.: Powierzchniowe przetwarzanie języka polskiego. Akademicka Oficyna Wydawnicza EXIT, Warsaw (2008)
23. Przepiórkowski, A., Buczyński, A.: ♠: Shallow Parsing and Disambiguation Engine. In: Vetulani, Z. (ed.) Proceedings of the 3rd Language & Technology Conference, Poznań, Poland, pp. 340–344 (2007)
24. Roland, D., Jurafsky, D.: How verb subcategorization frequencies are affected by corpus choice. In: Proceedings of COLING 1998, Montreal, pp. 1122–1128 (1998)
25. Schiehlen, M., Spranger, K.: Authomatic methods to supplement broad-coverage subcategorization lexicons. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, Lisbon, pp. 29–32. ELRA (2004)
26. Schulte im Walde, S.: Evaluating verb subcategorisation frames learned by a German statistical grammar against manual definitions in the *Duden* dictionary. In: Proceedings of the 10th EURALEX International Congress (2002)
27. Świdziński, M.: Syntactic dictionary of Polish verbs. Version 3a. Unpublished manuscript, University of Warsaw (1998)
28. Ushioda, A., Evans, D.A., Gibson, T., Waibel, A.: The automatic acquisition of frequencies of verb subcategorization frames from tagged corpora. In: Boguraev, B.K., Pustejovsky, J. (eds.) SIGLEX ACL Workshop of Acquisition of Lexical Knowledge from Text, Columbus, OH, pp. 95–106 (1993)
29. Vetulani, Z. (ed.): Proceedings of the 3rd Language & Technology Conference, Poznań, Poland (2007)
30. Woliński, M.: Komputerowa weryfikacja gramatyki Świdzińskiego. Ph. D. dissertation, Institute of Computer Science, Polish Academy of Sciences, Warsaw (2004)
31. Woliński, M.: An efficient implementation of a large grammar of Polish. Archives of Control Sciences 15(3), 251–258 (2005)