Semantyka a konfrontacja językowa, 3
SOW, Warszawa 2006

ADAM PRZEPIÓRKOWSKI
Instytut Podstaw Informatyki PAN
Warszawa

WHAT TO ACQUIRE FROM CORPORA IN AUTOMATIC VALENCE ACQUISITION[1]

1. Introduction

Recent years have witnessed an increased availability of linguistic corpora, which in turn has given rise to the blossoming of new and exciting research on the automatic learning of linguistic information implicitly contained in large collections of texts. One strand of this research, carried out since early 1990s (Webster and Marcus 1989, Brent 1991, 1993, Manning 1993), concerns the acquisition of valence information, i.e., information about subcategorisation (argument) frames of verbs and possibly other predicates. Crucially, syntactic valence information may be automatically acquired from corpora which are not annotated syntactically, e.g., from corpora annotated morphosyntactically: with grammatical classes (so-called parts of speech) and values of grammatical categories such as case and gender. Such a corpus is now available for Polish (`http://korpus.pl/`).

Most methods of valence acquisition follow a two-stage approach: first, various information is gleaned from corpora with the use of linguistic parsers and other text processing tools, and second, thus collected observations are subjected to statistical inference tests to decide which of them are significant and which are due to errors in corpora and text processing methods.

The aim of this article is to present the design of the kind and format of data that ought to be collected from corpora in the process of acquiring a valence dictionary of Polish.

[1]The work reported here has been conducted partially within the Ministry of Education and Science project number 3 T11C 003 28 and within the MNiSW/DAAD 2006–2007 cooperation project between the Institute of Computer Science PAS and the University of Regensburg.

1.1. Valence dictionaries for Polish

There are two relatively comprehensive published valence dictionaries for Polish, Polański 1992 and Bańko 2000, as well as an unpublished electronic valence dictionary, Świdziński 1998.[2] We discuss various design deficiencies of these dictionaries in Przepiórkowski 2003, including:

- lack of any formal criteria for distinguishing arguments and adjuncts,
- insufficient representation language which does not make it possible to represent relevant information about control/raising verbs, information about obligatory agreement between two arguments, etc.,
- purely morphosyntactic approach to valence.

The following subsection presents additional reasons for applying automatic methods to the development of valence dictionaries.

1.2. Why automatic valence acquisition?

As mentioned in Przepiórkowski and Fast 2005, which reports very preliminary experiments in valence acquisition for Polish, there are many well-known arguments for constructing valence dictionaries automatically, on the basis of naturally occurring texts:

> First of all, automatic methods of constructing valence dictionaries are much quicker and cheaper than the traditional manual process (e.g., the five volumes of Polański 1992 were published in the space of twelve years). Second, automatic methods are more objective than the traditional methods, based on potentially inconsistent intuitions of a team of lexicographers. Third, automatic methods may provide not only the categorical information, but also statistical information about how often a verb occurs with a given frame, which is particularly useful for probabilistic parsers. Fourth, the same methodology may be applied, without any overheads, to different collections of texts, e.g., to create thematic or diachronic valence dictionaries. Moreover, automatic methods may be and have been used for extending and verifying existing valence dictionaries.

Although it may seem that the results of such automatic methods must necessarily be of inferior quality, Schulte im Walde 2002 shows that the quality of such automatically constructed dictionaries may in fact be comparable to that of traditionally developed dictionaries. Currently, the main

[2] Apart from these, we are aware of three other much more limited resources of this kind: Morciniec *et al.* 1995, Mędak 2005 and a small electronic dictionary containing about 150 verbs created by Zygmunt Vetulani (Vetulani 2000). Another valence dictionary is in preparation; cf. Laskowski 2006.

disadvantage of automatic valence acquisition techniques is that they do not distinguish between different senses of a polysemous verb, but future advances in so-called Word Sense Disambiguation should to some extent alleviate this deficiency.

1.3. Basic assumptions

As postulated in Przepiórkowski 2004b, a valence dictionary should ideally be a syntactico-semantic lexicon, providing a semantic characteristics of deep (semantic) arguments of a verb and specifying how those arguments may be realised on the surface (morphosyntactically).

The aim of the current project is much more modest, namely, to automatically construct a syntactic valence dictionary, where only morphosyntactic information about arguments is provided, in the style of Świdziński 1998 and Bańko 2000. Also, just as in the former, no attempt is made at distinguishing between different senses of a verb.

Another design decision concerns the complement/adjunct distinction. Perhaps controversially, we do not assume any such distinction, for the simple reason that — after many decades of assuming such a dichotomy — linguists have not been able to propose an operational definition of this hypothetical distinction.[3] The consequence of this decision is that the resulting valence dictionary will in fact be a dictionary of dependents, rather than only arguments, co-occurring with verbs in Polish texts with significant frequency.

1.4. Two stages of valence acquisition

As mentioned above, there are usually two clearly separated stages in automatic valence acquisition. The first stage consists in collecting relevant information from corpora, usually information about co-occurrences of verbs and various phrases in the same sentence. For example, the sentence in (1) may give rise to the observation stated in (2), i.e., that a form of the verb DAĆ 'give' was observed occurring in a sentence together with a dative numeral phrase and an accusative adjectival phrase.

(1) *Daliśmy* *tym* *siedmiu* *marynarzom*
gave-PAST.PL.M1.PRI [these-DAT seven-DAT sailors-DAT]
najlepszą *z* *naszych* *łodzi.*
[best-ACC of our boats]
'We have given those seven sailors the best of our boats.'

(2) DAĆ 'give' <NumP[DAT], AdjP[ACC]>

[3] In Przepiórkowski 1999a, 1999b, 1999c, 2002, we discuss various tests aimed at distinguishing complements from adjuncts proposed in the linguistic literature and point out their vagueness and pairwise incompatibility.

The result of the first stage is a collection of such observations for all sentences occurring in a corpus. For example, the following information may be collected for the verb DAĆ:[4]

(3)

DAĆ	
23	<AdjP[ACC], NumP[DAT]>
385	<NP[ACC], NumP[DAT]>
987	<AdjP[ACC], NP[DAT]>
9242	<NP[ACC], NP[DAT]>
863	<NP[ACC], NP[DAT], NP[NOM]>
27	<NP[DAT], NP[GEN], NP[NOM]>
...	...

According to this table, forms of the verb DAĆ were observed in the context of an AdjP[ACC] and a NumP[DAT] 23 times, with a NP[ACC] and a NumP[DAT] — 385 times, etc. Such tables will be acquired for all verbs whose forms are heads of sentences in the corpus.

The results of this first stage of valence acquisition are usually noisy, that is, certain observations reflect errors in the corpus, or errors made at various stages of text processing. For example, it is possible that the phrase *obraz mojego brata* 'picture-ACC/NOM my-GEN brother-GEN', accusative in *Mojej siostrze dałem obraz mojego brata* 'To my sister, I gave a picture of my brother' is wrongly analysed as two phrases: *obraz* and *mojego brata*, with the first form wrongly interpreted as nominative rather than accusative. These errors would result in the observation <NP[DAT], NP[GEN], NP[NOM]> instead of the correct <NP[ACC], NP[DAT]>.

It is the task of the second stage to apply statistical tests to such data in order to decide which of these observations are reliable. At this stage the observations should also be generalised, e.g., it should be noticed that AdjP[ACC], NumP[ACC] and NP[ACC] are different realisations of the same accusative nominal position, etc.

The main aim of this article is to formally and precisely describe the format of the observations resulting from the first stage of valence acquisition. Note that the example above is deceptively simplistic. For example, the observations in (3) do not contain the information about the polarity of the verb, although it is a well-known fact that polarity affects valence requirements, the genitive of negation being the most conspicuous example of this class of phenomena. So, in order to distinguish genuine genitive-taking

[4]Such valence frames are treated as sets, i.e., the order of phrase specifications is not significant. The numbers below do not reflect any real corpus data.

verbs from accusative-taking verbs whose negated forms co-occur with genitive phrases, polarity information needs to be recorded.

On the other hand, a distinction is made in the examples above between nominal phrases and numeral phrases, which — as is well known — are distributionally (almost) fully equivalent; clearly, making such a distinction in the output of the first stage of valence acquisition needs to be justified. The following sections present and justify the design of the interface between the two stages of valence acquisition.

2. Verbal information

In (2) above, the only information recorded about the verb is its lemma, DAĆ. As noted in passing in the previous section, this information is not sufficient for drawing conclusions about possible valence frames of the verb; one other type of information that is necessary here is the polarity of the verbal form, i.e., whether it is negated or not.

2.1. Polarity

One reason for recording polarity[5] stems from the genitive of negation: if genitive phrases co-occur with negated forms of an otherwise accusative-taking verb, but hardly ever with non-negated forms of that verb, then very probably those genitive phrases are genitive-of-negation realisations of accusative valence requirements. On the other hand, if non-negated forms of a verb frequently occur with genitive phrases, then that verb probably has a valence frame with a genuine genitive argument.

Polarity information is also useful in distinguishing verbs such as POWIEDZIEĆ 'say', which may combine with sentential arguments headed either by the complementiser ŻE, or by the complementiser ŻEBY, from verbs such as WIERZYĆ 'believe', which normally combine with ŻE-sentences, but also allow ŻEBY-complements when negated. Finally, polarity will also be needed to distinguish inherently negative verbs such as NIE CIERPIEĆ 'detest, not stand' from their unmarked counterparts (here: CIERPIEĆ 'suffer').

2.2. Personal or impersonal?

Świdziński 1998 lists 5 verbs subcategorising for a non-subject nominative NP and 18 verbs combining with a non-subject nominative AdjP. Discovering such non-subject nominative arguments is difficult due to the fact that nominative subjects are often elided in Polish (so-called *pro*-drop), so the

[5]See Saloni and Świdziński 1998, pp. 156–161, Przepiórkowski 2000, and Kupść and Przepiórkowski 2002 for a more in-depth discussion of the issues mentioned in this subsection.

non-subject nominative argument may often be the only nominative phrase in a sentence and it might be incorrectly classified as the nominative subject.

For this reason it is important to record the information about the form of the verb, namely, whether it is personal or so-called *-no/-to* impersonal (e.g., *pito* 'drink', *jedzono* 'eat'). As is well known, such impersonal forms prohibit overt subjects, so any nominative phrase co-occurring with a *-no/-to* form of a verb may be treated as evidence that this verb takes a non-subject nominative argument, as in the following example, based on Saloni and Świdziński 1998, p. 128:

(4) *Wołano go Grubas.*
call-IMPS him-ACC Fat-NOM
'He was called Fat.'

Similarly, it makes sense to mark other verbal forms prohibiting nominative subjects, such as infinitive and imperative.

2.3. Agreement features

In order to identify non-subject nominative arguments, it is also useful to record agreement features of personal verbal forms: if the verbal form does not agree with a co-occurring nominative phrase, then this phrase may be a non-subject nominative argument of the verb, as in the example below:

(5) *Wołali go Grubas.*
call-PL.M1.TER him-ACC Fat-NOM.SG.M1
'They called him Fat.'

Another argument for recording agreement features of verbal forms will be provided in §4.3.

2.4. *się*

Finally, the results of the first stage of valence acquisition should include the information about the co-occurrence of verbal forms and the so-called reflexive marker *się*. In Polish, the reflexive marker is polyfunctional: it may be a part of an inherently reflexive verb (e.g., ŚMIAĆ SIĘ 'laugh'), it may mark an impersonal or middle construction, and perhaps it may also be an anaphoric (reflexive or reciprocal) realisation of an argument. Moreover, as discussed in Kupść (1999), a single occurrence of *się* may simultaneously play a number of roles.

Because of this polyfunctionality of *się*, and since impersonal and middle uses of *się* influence the observed valence (impersonal prohibits the subject, middle additionally promotes the object), its occurrences should only be recorded in the first stage of valence acquisition and used in statistical and

linguistic reasoning in the second stage to distinguish between inherently reflexive verbs on one hand and impersonal, middle or anaphoric uses with non-reflexive verbs, on the other. In practice, we will treat *się* as a separate argument type (see (6b) and §3 below).

2.5. Summary

Following the decisions of the previous subsections, (1) above (p. **??**) should lead to observation (2′), while the sentence (6a) should give rise to (6b).

(2′) DAĆ:AFF:FIN.PL.M1.PRI <NumP[DAT], AdjP[ACC]>

(6a)

Ta	*książka*	*nie*	*sprzedawała*	*się*	*dobrze.*
this-NOM	book-NOM	NEG	sell-PAST.SG.F.TER	RM	well

'This book didn't sell well.'

(6b) SPRZEDAWAĆ:NEG:FIN.SG.F.TER <NP[NOM], SIĘ, AdvP>

In general, the information recorded for each verb will consist of:

- the lemma (e.g., DAĆ or SPRZEDAWAĆ),
- the polarity (AFF or NEG),
- the morphosyntactic class (FIN — personal form, IMPT — imperative form, INF — infinitival form, IMPS — impersonal form),[6]
- in case of personal forms, also number, gender and person.

3. Grammatical classes

Following and extending Świdziński 1998, Saloni and Świdziński 1998 and Polański 1992, we assume the repertoire of basic argument types listed below:

(7)

NP	nominal phrase
NumP	numeral phrase
AdjP	adjectival phrase
PrepNP	prepositional-nominal phrase
PrepNumP	prepositional-numeral phrase
PrepAdjP	prepositional-adjectival phrase
AdvP	adverbial phrase
InfP	infinitival phrase
CP	sentential phrase introduced by a complementiser
KP	sentential phrase introduced by an interrogative or relative phrase
OR	oratio recta (direct speech)
się	reflexive marker

[6] Participial and gerundive forms will be treated as forming adjectival, adverbial or nominal phrases.

Argument types NP, AdjP, PrepNP, PrepAdjP, InfP, CP and OR are uncontroversial and can be found in all three Polish valence dictionaries mentioned in §1.1.[7]

Two argument types absent in the works cited above are NumP and PrepNumP, for numeral phrases and prepositional phrases with numeral immediate constituents. Their separate treatment will make it possible to find out whether there exist Polish verbs significantly preferring or maybe even strictly subcategorising for numeral phrases with the exclusion of standard nominal phrases. Such a possibility cannot be *a priori* ruled out as, contrary to the common assumption that the distribution of nominal phrases and numeral phrases is identical, there exist contexts making the distinction. The most interesting such a case is perhaps the distributional preposition PO, which combines with locative nominal phrases but accusative numeral phrases (Łojasiewicz, 1979, Przepiórkowski, 2006). Also within complex numeral phrases particular positions may be filled by numerals, but not nouns (Gruszczyński and Saloni, 1978).

Unlike in Saloni and Świdziński, 1998, and Świdziński, 1998, who allow prepositional realisations of adverbial phrases, AdvP is understood here purely morphosyntactically, as a phrase headed by an adverb or an adverbial participle. Moreover, we treat the reflexive marker *się* as a separate argument type, on the basis of considerations in § 2.4.

Finally, it is difficult to automatically distinguish embedded questions, true relative clauses and free relatives. As the examples below show, the clause *kto przybył pierwszy* may be involved in all three constructions.

(8a) *Zapytał mnie wczoraj ktoś, kto przybył pierwszy.*
asked me yesterday somebody who arrived first
'Somebody asked me yesterday who arrived first.'
'A person that arrived first asked me yesterday.' (ellipsis)

(8b) *Nagrodę wygrał ktoś, kto przybył pierwszy.*
prize-ACC won somebody-NOM who arrived first
'The person that arrived first won the prize.'

(8c) *Nagrodę wygrał, kto przybył pierwszy.*
prize-ACC won who arrived first
'Whoever arrived first, won the prize.'

For this reason, the first stage of valence acquisition will only register

[7] Of course, the notation of these dictionaries differs substantially. Moreover, although Polański 1992 does not mention adjectival arguments in the introduction, Adj is actually used in lexical entries; cf., e.g., entries for BYĆ and UWAŻAĆ.

the presence of KP, i.e., a clause starting with an interrogative or relative phrase,[8] with the interpretation of such data being left for the second stage.

4. Grammatical categories

4.1. Case

All valence dictionaries specify the case of nominal and adjectival arguments and of nominal and adjectival phrases immediately within prepositional arguments; we have also done so in the examples (e.g., (2) and (6b)). In terms of the basic argument types established in the previous subsection, that means that the following argument types will be specified for case: NP, NumP, AdjP, PrepNP, PrepNumP and PrepAdjP.

4.2. Aspect

Although no other purely morphosyntactic information is usually specified in valence dictionaries, it is well known (Saloni and Świdziński, 1998, p. 137) that there exist verbs subcategorising for infinitival phrases with of a specific aspect: imperfective (e.g., ZACZĄĆ 'start', PRZESTAĆ 'stop' and other phasal verbs) or perfective (ZDOŁAĆ 'manage', ZDĄŻYĆ 'manage on time'). In order to automatically discover such facts, it is necessary to record aspect values of InfP arguments.

4.3. Number, gender, person

It turns out that it also makes sense to record so-called ϕ-features (number, gender, person) of certain phrases, mainly for the purpose of discovering obligatory number and gender agreement between two arguments, as in the case of the verbs BYĆ 'be (copula)' or WYGLĄDAĆ 'seem':

(9) *Ta* *lampa* *jest* *piękna.*
this lamp-SG.NOM.F is beautiful-SG.NOM.F
'This lamp is beautiful.'

(10) *Ta* *lampa* *wygląda* *na* *piękną.*
this lamp-SG.NOM.F seems PREP beautiful-SG.ACC.F
'This lamp seems beautiful.'

Such agreement involves a 'controlling' nominative, numeral or adjectival phrase and a 'controlled' adjectival (cf. (9)) or prepositional-adjectival (cf. (10)) phrase, so number and gender information should be recorded for NP, NumP, AdjP and PrepAdjP.

[8] Note that, strictly speaking, KPs do not need to begin with an interrogative or relative word, e.g.: *(Zastanawiam się,) na kogo czeka* '(I wonder) for whom (s)he is waiting'; hence our loose use of the term 'interrogative or relative *phrase*'.

Note that when the controlling argument is the subject, the evidence for the agreement may be indirect: in case the controlling subject is *pro*-dropped, its ϕ-features are reflected only on the verb, so the observed agreement between the verb and an adjectival phrase may be treated as some evidence for the agreement between the subject and that adjectival phrase. This is the additional reason for recording agreement features of the verb that we alluded to in §2.3. above.

Finally, one more type of clue for detecting non-subject nominative arguments may be provided by the value of person of nominative phrases: if it does not agree with the person value of the verb, then that nominative phrase may be a non-subject nominative argument.

4.4. Summary

The table below summarises the considerations of the previous subsections.

(11)

grammatical category	***values recorded for***
case	NP, NumP, AdjP, PrepNP, PrepNumP, PrepAdjP
aspect	InfP
number, gender	NP, NumP, AdjP, PrepAdjP
person	NP[NOM]

5. Lexical information

Valence dictionaries usually report only a very limited lexical information about the argument, mainly the preposition of a prepositional argument and the complementiser of a sentential argument. However, Polański 1992 and Bańko 2000 also give some information about the semantic class of an argument; for example, in Polański 1992 this information ranges from the general [±Human], [±Animate] and [±Abstract] features, to more specific information such as [Institution], [Liquid] or [Machine].

Although acquiring such selectional restrictions is not in the immediate scope of the current project, we would like to record the information needed to discover such restrictions for future work. For this reason, it is necessary to also record the lemma of the main contentful form of each argument. Note that this means that in case of some argument types two lemmata need to be recorded: the syntactic head, e.g., the preposition, and the semantic head, e.g., the noun heading the nominal phrase within the prepositional phrase. Similarly, in case of a sentential argument, both the complementiser (syntactic head) and the main verb of the subordinate sentence (semantic head) should be recorded. By analogy, in case of KP arguments, we will

treat the initial interrogative or relative word as the syntactic head and the main verb as the semantic head. The distinction between the syntactic head and the semantic head will also be made for numeral phrases.

On the other hand, in case of NP, AdjP, AdvP and InfP, the syntactic head will be the same as the semantic head. Finally, we will assume that OR and *się* arguments are headless.

The table below summarises these considerations.

(12)

argument type	***syntactic head***	***semantic head***
NP	noun	
NumP	numeral	semantic head of the sister of the numeral
AdjP	adjective	
PrepNP, PrepNumP, PrepAdjP	preposition	semantic head of the sister of the preposition
AdvP	adverb	
InfP	infinitive verb	
CP	complementiser	main verb
KP	interrogative or relative word	main verb
OR	—	
się	—	

6. Summary and BNF grammar

On the basis of the considerations above, we propose the result of the first stage of valence acquisition to be a list of lexeme observations, where each lexeme observation consists of a lemma (cf. ZAORDYNOWAĆ 'order' below) and a number of frame observations. Each frame observation consists of an integer specifying how many times this observation has been made, verb information and frame information, where verb information specifies the polarity and the morphosyntactic interpretation of the form of the given lexeme, while the frame information is a sequence of argument specifications. Finally, each argument specification consists of the grammatical class information, the syntactic head (if applicable), the semantic head (if different from the syntactic head) and the relevant morphosyntactic information.

(13) presents, in the format used in the previous sections, the full lexeme information for ZAORDYNOWAĆ 'order' which might be gleaned from the IPI PAN Corpus.[9]

[9]The frame observations in (13) correspond to the correct disambiguations of the rel-

(13) ZAORDYNOWAĆ

1	AFF:FIN.SG.M1.TER	<>
1	AFF:FIN.SG.M1.TER	<OR>
1	AFF:FIN.SG.M1.TER	<NP[PROFESOR,NOM.SG.M1.TER], NP[AMERYKA,ACC.SG.F]>
1	AFF:FIN.SG.M1.TER	<OR, NP[PAN,NOM.SG.M1.TER], AdvP[ZWRACAĆ]>
1	AFF:FIN.SG.M1.TER	<OR, AdvP[RZECZOWO]>
1	AFF:FIN.SG.M1.TER	<PP[PO,PRZYLOT,LOC], AdjP[PODOPIECZNY,DAT.PL.M1], NP[TRENING,ACC.SG.M3]>
1	AFF:IMPS	<NP[SKRĘT,ACC.PL.M3]>
1	AFF:FIN.SG.M1.TER	<PP[PRZEZ,TYDZIEŃ,ACC], NP[TRENER,NOM.SG.M1.TER], NP[ON,DAT.PL.F], NP[PORCJA,ACC.SG.F]>

A complete Backus-Naur Form grammar for the actual computer-readable format of lexeme observations is presented below. Non-terminals ⟨lem⟩ and ⟨form⟩ represent sequences of letters corresponding to lemmata and forms (the latter for interrogative or relative pronouns; cf. ⟨kp⟩), while ⟨integer⟩ represents a positive integer number. Possible realisations of ⟨number⟩, ⟨case⟩, ⟨gender⟩, ⟨person⟩ and ⟨aspect⟩ represent the values of corresponding morphosyntactic categories used in the IPI PAN Corpus (cf. Przepiórkowski 2004a for details).

(14) ⟨lexeme_observation⟩ ::= ⟨lem⟩ ⟨frame_observations⟩
⟨frame_observations⟩ ::= ⟨frame_observation⟩
| ⟨frame_observation⟩ ⟨frame_observations⟩
⟨frame_observation⟩ ::= ⟨integer⟩ ⟨polarity⟩`:`⟨morph⟩ ⟨frame⟩
⟨polarity⟩ ::= `aff` | `neg`
⟨morph⟩ ::= `fin.`⟨number⟩`.`⟨gender⟩`.`⟨person⟩
| `impt.`⟨number⟩`.`⟨person⟩
| `inf`
| `imps`
⟨frame⟩ ::= `<>`
| `<`⟨arguments⟩`>`

evant results of the query `[base=zaordynować]` on the 30-million sample of the IPI PAN Corpus (2nd edition; so-called `2.sample.30`, cf. `korpus.pl`; results number 2–6, 8, 9, 11).

⟨arguments⟩ ::= ⟨argument⟩
 | ⟨argument⟩`,` ⟨arguments⟩
⟨argument⟩ ::= ⟨np⟩ | ⟨nump⟩ | ⟨adjp⟩ | ⟨prepnp⟩ | ⟨prepnump⟩
 | ⟨prepadjp⟩ | ⟨advp⟩ | ⟨infp⟩ | ⟨cp⟩ | ⟨kp⟩ | ⟨or⟩ | ⟨sie⟩
⟨np⟩ ::= `np:`⟨lem⟩`:`⟨number⟩`.`⟨case_non_nom⟩`.`⟨gender⟩
 | `np:`⟨lem⟩`:`⟨number⟩`.`⟨case_nom⟩`.`⟨gender⟩`.`⟨person⟩
⟨nump⟩ ::= `nump:`⟨lem⟩`:`⟨lem⟩`:`⟨number⟩`.`⟨case⟩`.`⟨gender⟩
⟨adjp⟩ ::= `adjp:`⟨lem⟩`:`⟨number⟩`.`⟨case⟩`.`⟨gender⟩
⟨prepnp⟩ ::= `prepnp:`⟨lem⟩`:`⟨lem⟩`:`⟨case⟩
⟨prepnump⟩ ::= `prepnump:`⟨lem⟩`:`⟨lem⟩`:`⟨case⟩
⟨prepadjp⟩ ::= `prepadjp:`⟨lem⟩`:`⟨lem⟩`:`⟨number⟩`.`⟨case⟩`.`⟨gender⟩
⟨advp⟩ ::= `advp:`⟨lem⟩
⟨infp⟩ ::= `infp:`⟨lem⟩`:`⟨aspect⟩
⟨cp⟩ ::= `cp:`⟨lem⟩`:`⟨lem⟩
⟨kp⟩ ::= `kp:`⟨form⟩`:`⟨lem⟩
⟨or⟩ ::= `or`
⟨sie⟩ ::= `sie`
⟨number⟩ ::= `sg` | `pl`
⟨case⟩ ::= ⟨case_nom⟩ | ⟨case_non_nom⟩
⟨case_nom⟩ ::= `nom`
⟨case_non_nom⟩ ::= `gen` | `dat` | `acc` | `inst` | `loc` | `voc`
⟨gender⟩ ::= `m1` | `m2` | `m3` | `f` | `n`
⟨person⟩ ::= `pri` | `sec` | `ter`
⟨aspect⟩ ::= `imperf` | `perf`

According to this grammar, the information in (13) will be represented as in (13′) below (white spaces are not significant).

(13′)
```
zaordynować 1 aff:fin.sg.m1.ter <> 1 aff:fin.sg.m1.ter
<or> 1 aff:fin.sg.m1.ter <np:profesor:nom.sg.m1.ter,
  np:ameryka:acc.sg.f>
1 aff:fin.sg.m1.ter <or, np:pan:nom.sg.m1.ter,
  advp:zwracać>
1 aff:fin.sg.m1.ter <or, advp:rzeczowo>
1 aff:fin.sg.m1.ter <prepnp:po:przylot:loc,
  adjp:podopieczny:dat.pl.m1, np:trening:acc.sg.m3>
1 aff:imps <np:skręt:acc.pl.m3> 1 aff:fin.sg.m1.ter
<prepnp:przez:tydzień:acc,
  np:trener:nom.sg.m1.ter, np:on:dat.pl.f,
  np:porcja:acc.sg.f>
```

7. Concluding remarks

All literature on automatic valence acquisition that we are familiar with, including Brent 1993, Manning 1993, Briscoe and Carroll 1997, Korhonen 2002 and Schulte im Walde 2002, as well also our earlier work (Przepiórkowski and Fast 2005, Fast and Przepiórkowski 2005), concentrates on the second, statistical stage of valence acquistion, taking for granted both the repertoire of valence information (tagset, phrase types) and means of extracting relevant information from corpora (parsers). This is caused by practical reasons: on the one hand, designing the desired output of the first stage of valence acquisition and developing corresponding tools able to produce such an output is a time-consuming task requiring some linguistic insight, and, on the other, various parsers are available for languages that have been the focus of valence acquisition research so far, i.e., for English and German. However, often the outcome of such off-the-shelf tools is not ideal for valence acquisition.

The context of the research reported here is different: it is based on a corpus developed within a project led by the author (Przepiórkowski 2004a), annotated with a tagset co-designed by the author (Przepiórkowski and Woliński 2003a, 2003b), and the tools used for collecting relevant information from corpora are being constructed by the author,[10] so there is a unique opportunity to homogeneously design and control the whole process of valence acquisition.

Let us, however, end on a cautionary note: the above formal specification of the morphosyntactic information collected from corpora for the purpose of automatic valence acquistion, although more detailed, more extensive and perhaps more carefully designed than in case of similar valence acquisition attempts for other languages, has been developed mostly *a priori*, and its usefulness and correctness still needs to be verified in practice. We hope to be able to report on the outcome of this verification in the near future.

References

Bańko 2000: Bańko, M., editor., *Inny słownik języka polskiego*. Wydawnictwo Naukowe PWN, Warsaw.

Borsley and Przepiórkowski 1999: Borsley, R. D. and Przepiórkowski, A., editors., *Slavic in Head-Driven Phrase Structure Grammar*. CSLI Publications, Stanford, CA.

[10] Possibly, a deep parser developed at the author's host institute will also be used; cf. Woliński 2004.

Brent 1991: Brent, M. R., Automatic semantic classification of verbs from their syntactic contexts. In *Proceedings of the European Association for Computational Linguistics (EACL) 1991, Berlin, Germany, 9–11 April 1991*, pages 222–226. EACL.

Brent 1993: Brent, M. R., From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics*, 19(2), 243–262.

Briscoe and Carroll 1997: Briscoe, T. and Carroll, J., Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, pages 356–363.

Fast and Przepiórkowski 2005: Fast, J. and Przepiórkowski, A., Automatic extraction of Polish verb subcategorization: An evaluation of common statistics. In Z. Vetulani, editor, *Proceedings of the* 2nd Language & Technology Conference, pages 191–195, Poznań, Poland.

Gruszczyński and Saloni 1978: Gruszczyński, W. and Saloni, Z., Składnia grup liczebnikowych we współczesnym języku polskim. *Studia Gramatyczne*, II, 17–42.

Korhonen 2002: Korhonen, A., *Subcategorization Acquisition*. Ph. D. dissertation, University of Cambridge.

Kosta and Frasek 2002: Kosta, P. and Frasek, J., editors., *Current Approaches to Formal Slavic Linguistics*, Frankfurt am Main. Peter Lang.

Kupść 1999: Kupść, A., Haplology of the Polish reflexive marker. In Borsley and Przepiórkowski(1999), pages 91–124.

Kupść and Przepiórkowski 2002: Kupść, A. and Przepiórkowski, A., Morphological aspects of verbal negation in Polish. In Kosta and Frasek(2002), pages 337–346.

Laskowski 2006: Laskowski, R., Słownik łączliwości czasowników (kilka podstawowych zasad). This volume.

Łojasiewicz 1979: Łojasiewicz, A., O budowie wyrażeń z przyimkiem *po* dystrybutywnym. *Polonica*, **V**, 153–160.

Manning 1993: Manning, C. D., Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st ACL*, pages 235–242.

Mędak 2005: Mędak, S., *Praktyczny słownik łączliwości składniowej czasowników polskich*. Universitas, Cracow.

Morciniec *et al.* 1995: Morciniec, N., Cirko, L., and Ziobro, R., *Słownik walencyjny czasowników niemieckich i polskich / Wörterbuch zur Valenz Deutscher und Polnischer Verben*. Wydawnictwo Uniwersytetu Wrocławskiego, Wrocław.

Polański 1992: Polański, K., editor, *Słownik syntaktyczno-generatywny czasowników polskich*. Zakład Narodowy im. Ossolińskich / Instytut Języka Polskiego PAN, Wrocław / Cracow (1980–1992).

Przepiórkowski 1999a: Przepiórkowski, A., *Case Assignment and the Complement-Adjunct Dichotomy: A Non-Configurational Constraint-Based Approach*. Ph. D. dissertation, Universität Tübingen.

Przepiórkowski 1999b: Przepiórkowski, A., On case assignment and 'adjuncts as complements'. In G. Webelhuth, J.-P. Koenig, and A. Kathol, editors, *Lexical and Constructional Aspects of Linguistic Explanation*, pages 231–245. CSLI Publications, Stanford, CA.

Przepiórkowski 1999c: Przepiórkowski, A., On complements and adjuncts in Polish. In Borsley and Przepiórkowski(1999), pages 183–210.

Przepiórkowski 2000: Przepiórkowski, A., Long distance genitive of negation in Polish. *Journal of Slavic Linguistics*, **8**, 151–189.

Przepiórkowski 2002: Przepiórkowski, A., Verbal proforms and the structural complement-adjunct distinction in Polish. In Kosta and Frasek(2002), pages 405–414.

Przepiórkowski 2003: Przepiórkowski, A., On the computational usability of valence dictionaries for Polish. IPI PAN Research Report 971, Institute of Computer Science, Polish Academy of Sciences.

Przepiórkowski 2004a: Przepiórkowski, A., *The IPI PAN Corpus: Preliminary version*. Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Przepiórkowski 2004b: Przepiórkowski, A., Towards the design of a syntactico-semantic lexicon for Polish. In M. A. Kłopotek, S. T. Wierzchoń, and K. Trojanowski, editors, *Intelligent Information Processing and Web Mining*, Advances in Soft Computing, pages 237–246. Springer-Verlag, Berlin.

Przepiórkowski 2006: Przepiórkowski, A., O dystrybutywnym PO i liczebnikach jedynkowych. Unpublished manuscript.

Przepiórkowski and Fast 2005: Przepiórkowski, A. and Fast, J., Baseline experiments in the extraction of Polish valence frames. In M. A. Kłopotek, S. T. Wierzchoń, and K. Trojanowski, editors, *Intelligent Information Processing and Web Mining*, Advances in Soft Computing, pages 511–520. Springer-Verlag, Berlin.

Przepiórkowski and Woliński 2003a: Przepiórkowski, A. and Woliński, M., A flexemic tagset for Polish. In *Proceedings of* Morphological Processing of Slavic Languages*, EACL 2003*, pages 33–40, Budapest.

Przepiórkowski and Woliński 2003b: Przepiórkowski, A. and Woliński, M., The unbearable lightness of tagging: A case study in morphosyntactic tagging of Polish. In *Proceedings of the* 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*, EACL 2003*, pages 109–116.

Saloni and Świdziński 1998: Saloni, Z. and Świdziński, M., *Składnia współczesnego języka polskiego*. Wydawnictwo Naukowe PWN, Warsaw, 4th (changed) edition.

Schulte im Walde 2002: 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.*

Schulte im Walde 2003: Schulte im Walde, S., *Experiments on the Automatic Induction of German Semantic Verb Classes.* Ph.D. thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Stuttgart.

Świdziński 1998: Świdziński, M., Syntactic dictionary of Polish verbs. Unpublished manuscript, Version 3a, Warsaw University.

Vetulani 2000: Vetulani, Z., Electronic language resources for Polish: POLEX, CEGLEX and GRAMLEX. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, and G. Stainhaouer, editors, *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000*, pages 367–374. ELRA.

Webster and Marcus 1989: Webster, M. and Marcus, M., Automatic acquisition of the lexical frames of verbs from sentence frames. In *Proceedings of the 27th Meeting of the ACL*, pages 177–184.

Woliński 2004: Woliński, M., *Komputerowa weryfikacja gramatyki Świdzińskiego.* Ph. D. dissertation, Institute of Computer Science, Polish Academy of Sciences, Warsaw.