On Heads and Coordination in Valence Acquisition

Adam Przepiórkowski

Institute of Computer Science
Polish Academy of Sciences,
Warsaw, Poland
adamp@ipipan.waw.pl
http://nlp.ipipan.waw.pl/~adamp/

Abstract. The aim of this paper is to present the design of a partial syntactic annotation of the IPI PAN Corpus of Polish [22] and the corresponding extension of the corpus search engine Poliqarp [25,12] developed at the Institue of Computer Science PAS and currently employed in Polish and Portuguese corpora projects. In particular, we will argue for the need to distinguish between, and represent both, syntactic and semantic heads, and we will sketch the representation of coordination, the area traditionally controversial both in theoretical and in computational linguistics. The annotation is designed in a way intended to maximise the usefulness of the resulting corpus for the task of automatic valence acquisition.

1 Introduction

1.1 Motivation and Outline

Treebanks are resources often used for the automatic acquisition of linguistic and natural language processing (NLP) knowledge such as frequencies of particular constructions or phrase types, syntactic valence or collocational information.[1]

The aim of this article is to present the design of a treebank to be used specifically for the purposes of automatic valence acquisition,[2] where both morphosyntactic and lexico-semantic selectional requirements will be learned. For this reason, it is necessary to identify both the syntactic head (for morphosyntactic valence constraints) and the semantic head (for lexico-semantic selectional restrictions) of any construction. Section 2 shows that semantic heads cannot be deduced automatically from the syntactic structure. But if both the syntactic and the semantic head are annotated for any constructions, then the unsolved question of the headedness of coordinate structures becomes even more pressing; a possible solution is proposed in section 3.

The treebank mentioned above will be built in two stages. First, a partial treebank will be constructed with the help of shallow grammars which will identify NPs, PPs, and other possible verbal dependents. No attempt will be made

[1] This article is an extended and corrected version of [23].

[2] But, no doubt, this resource will turn out to be useful also for many other purposes.

A. Gelbukh (Ed.): CICLing 2007, LNCS 4394, pp. 50–61, 2007.

at constructing the full structure of a clause at this stage. That resulting information will be used to automatically construct a preliminary valence dictionary (cf. [24] and [7] for recent relevant experiments on Polish). The second stage will consist in the manual construction of full parses for clauses, possibly on the basis of the results of automatic deep parsing (with the use of the valence dictionary created in the first stage).

This paper reports on work within the first stage. After discussing the syntactic/semantic head distinction in §2 and coordination in §3, we propose an XML representation for such annotation in §4 and, in §5, we describe a conservative extension of the query language used by the Poliqarp search engine that takes advantage of such grammatical annotation. The remainder of this section briefly presents the current status of the IPI PAN Corpus of Polish, which constitutes the empirical basis for the planned treebank, and Poliqarp, the search tool used to query the corpus.

1.2 The IPI PAN Corpus

The IPI PAN Corpus of Polish ([22]; `http://korpus.pl/`), presently the only morphosyntactically annotated large corpus of Polish, was first made available for search in June 2004. The whole corpus contains over 250 million segments (over 200 million orthographic words; punctuation marks count as separate segments and some orthographic words are split into smaller segments for good linguistic reasons described in [26,27]). Recently, a source (XML) version of a subcorpus containing 100 million segments has been made available to the public for non-commercial research purposes.[3] Unique features of this corpus include a carefully designed and documented morphosyntactic tagset and the inclusion of all possible morphosyntactic interpretations, in addition to those chosen by the tagger as correct in the given context. The corpus is XML-encoded according to (slightly modified) XCES [11] specifications.

1.3 Poliqarp

Poliqarp is an indexing and searching tool developed in the same project as the IPI PAN Corpus, but it was designed as a universal corpus management tool: the tagset may be specified externally and the internal character coding is UTF-8, so the tool could be used for any corpus of any language.[4] A stable version 1.0 of Poliqarp was made available to the community under the GNU GPL licence (cf. `http://korpus.pl/index.php?page=poliqarp`).

The syntax query of Poliqarp is based on that of CQP [5], but it contains some unique features. One of the most interesting is that one may refer both to all morphosyntactic interpretations given by the morphological analyser and to the disambiguated interpretations; for example, the query '`[case~acc & case=gen]`' may be used to find those forms which were tagged as genitive but which may, in

[3] See `http://korpus.pl/index.php?page=download` for details.

[4] It has been used recently for the Portuguese corpus developed by António Branco's group in Lisbon, [1], cf. `http://lxcorpus.di.fc.ul.pt/`.

other contexts, be analysed as accusative. Moreover, since some contexts do not provide sufficient information to fully disambiguate a form, Poliqarp allows to distinguish between certain and uncertain information. For example, the query '`[case=gen]`' may be used to search for any forms whose disambiguated interpretation (possibly one of many, if the tagger could not narrow down interpretations to one) is genitive, while '`[case==gen]`' finds those forms that have a unique (certain) genitive interpretation.[5]

See [22] for a detailed description of the tagset and the query language.

2 Distinguishing Syntactic and Semantic Heads

It is well known that valence must be expressed both at the syntactic and at the semantic level; a verb (or any predicate) may refer to the morphosyntactic (e.g., part of speech, case) or the lexico-semantic (e.g., volition, humanness) properties of its argument. For this reason, both the syntactic head and the semantic head of a potential dependent must be made available to the valence acquisition algorithm.[6]

In many cases syntactic heads are also semantic heads, as in the majority of noun phrases, but there are exceptions. In many cases, the syntactic structure of a construction allows one to automatically deduce the semantic head, as in the case of the English determiner+noun NPs, where the noun is always the semantic head, although the determiner may be taken to be the syntactic head, but again there are exceptions. For these reasons it is necessary to explicitly represent both the syntactic head and the semantic head in a treebank.

One area where it is very difficult to automatically recognise the semantic head on the basis of syntax only is the domain of numeral and nominal phrases in Polish. In Polish, numerals are a morphosyntactic rather than a semantic class; when in subject position, they exhibit a special agreement pattern with the verb, which occurs in the 'default' 3rd person singular neuter form rather than in the form which would agree with the noun. For example 'Five books lay on the table' would be translated into *Pięć książek*-GEN.FEM.PL *leżało*-3RD.NEUT.SG *na stole* (lit.: 'five books lay on table') rather than **Pięć książek*-GEN.FEM.PL *leżały*-3RD.FEM.PL *na stole*. It is commonly assumed that numerals are the syntactic

[5] Let us mention, for completeness, that the query '`[case~~gen]`' would find all forms which are unambiguously genitive, regardless of context, i.e., forms whose all possible interpretations are genitive.

[6] Note that this distinction is understood here, roughly, as approximating Mel'čuk's distinction between the morphological and syntactic dependency on one hand, and the semantic dependency on the other hand (cf. [16] for a summary and references), rather than as in Head-driven Phrase Structure Grammar (HPSG) [19,20], where so-called adjuncts are always semantic heads because they are semantic functors. The notion of *semantic head* corresponds to the notion of "useful head" in [31] or "lexical head" used interchangably with "semantic head" in [8]. A distinction between syntactic heads and semantic heads was already known by the modistic grammarian Radulphus Brito (c. 1300), cf. [6].

heads of such numeral constructions [29,21], while the nouns are semantic heads. On the other hand, the noun is both the syntactic and the semantic head in a noun phrase. However, there are number-denoting lexemes which are clear morphosyntactic nouns, e.g., *tuzin* 'dozen', where it is the complement of the syntactic head noun that should be analysed as the semantic head, and there are also lexemes such as *tysiąc* 'thousand' which are morphosyntactically ambigous between the numeral and the nominal interpretation. In fact, various measure phrases are widely discussed cases of the syntactic/semantic head mismatch in various languages, cf., e.g., [35] for English and [31] for French, with a broader spectrum of such mismatches in nominal phrases, involving phrases like *part of the room*, *herd of wildebeest*, *kind of fish*, *bout of the flu* and *her jerk of a husband*, discussed in [8] and [9].

Another area where syntax does not pre-determine semantic headedness are adjectival phrases: there is a subset of (syntactically) adjectival phrases, called elective phrases, as in *największy z chłopców* '(the) biggest of boys', where the semantic head is actually the noun argument of the proposition *z* 'of' subcategorised for by the comparative or superlative form of the adjective (*największy* 'biggest' in this example).

More examples can be given of Polish constructions whose purely morphosyntactic makeup does not determine the semantic headedness. For this reason, if a treebank is to be useful in applications such as exhaustive valence extraction, it must explicitly encode both kinds of headedness.

3 Coordination

Coordination is one of the most controversial areas in theoretical linguistics. In particular, it is far from clear what should count as the head in coordinate constructions. Postulating the existence of two possibly different heads makes things even worse: while many syntactic theories take the conjunction to be the syntactic head, it clearly is not the semantic head. In fact, each conjunct should be treated as a semantic head.

This is exactly the stance that we adopt here: since — assuming that a coordinated structure has a semantic head — all conjuncts should be treated as heads, we will assume that coordinations are actually multi-headed structures, with each conjunct providing a syntactic head and a corresponding semantic head.

This decision is also dictated by valence acquisition considerations: in cases of coordination of unlike categories [28], the coordinate structure provides evidence for two syntactically different valence frames of the same verb. For example, the sentence *Opowiadał o Wenecji i że musi tam wrócić* '(He) was saying about Venice and that (he) must return there' (from [13]) is grammatical only because the verb OPOWIADAĆ ('talk', 'say') may be combined either with a prepositional phrase headed by the preposition O or with a clause headed by the complementiser ŻE. This evidence would be missed, or at least it would have to be

reached via much more complicated reasoning, if the conjunction or just one of the conjuncts were taken to be the syntactic head.[7]

Note that this treatment of coordination makes coordinate structures essentially multi-headed, as in [3] (or, in a way, as in [33] and in a mediaeval modistic grammar [6], where a coordinate structure is not a phrase in its own right, but the verb has a direct relation to each of the conjuncts), unlike in modern linguistic theories, which often analyse coordination as head-argument constructions, either by postuling that coordinate constructions are headed by the first conjunct (e.g., [16]), or that they are headed by conjunction (e.g., [30]). We believe that the cases of coordination of unlike categories, such as mentioned above, while providing practical reasons for the treatment of coordinate structures as multi-headed in the context of a valence acquisition project, also constitute a strong evidence for such a multi-headed theoretical linguistic analysis of coordination.[8]

The final argument for this treatment of coordination comes from the design of the query syntax to be discussed in §5.

4 XML Representation

Each text in the IPI PAN Corpus of Polish currently consists of three XML files: `header.xml`, containing metadata, `text.xml`, validated by the (slightly modified) `xcesDoc.dtd` from the XCES (XML Corpus Encoding Standard; [11]) specification, containing the text itself with some structural annotation, and `morph.xml`, validated by the (slightly modified) `xcesAna.dtd`, containing morphosyntactic annotation.

Each `morph.xml` is sequence of `<tok>` elements grouped into sentences (`<chunk type="s">` elements), which are in turn grouped into paragraphs (`<chunk type="p">` elements). A three-segment fragment of a `morph.xml`, translated as 'for (the) Częstochowa steel-mill', is given below:[9]

```
<tok id="tA10">
<orth>dla</orth>
<lex disamb="1"><base>dla</base><ctag>prep:gen</ctag></lex>
</tok>
```

[7] It should be noted that the coordination of unlike categories is systematically (if not textually) common in Polish, e.g., [32,13] discusses various other cases of coordination involving an NP and a clause, [14] discusses many cases of coordination involving an NP and a PP, [21] gives examples of coordination of NPs of different cases, etc.

[8] An alternative theory that can easily account for such data is an ellipsis-based theory of (apparent) non-constituent coordination of [2]. In general, HPSG is perhaps unique among contemporary theories in directly addressing various difficult problems of coordinate structures and proposing explicit solutions.

[9] As mentioned above, all morphosyntactic interpretations are retained for each segment, but the one that the tagger ruled as correct is marked with the '`disamb="1"`' attribute.

```
<tok id="tA11">
<orth>Huty</orth>
<lex disamb="1"><base>huta</base><ctag>subst:sg:gen:f</ctag></lex>
<lex><base>huta</base><ctag>subst:pl:nom:f</ctag></lex>
<lex><base>huta</base><ctag>subst:pl:acc:f</ctag></lex>
<lex><base>huta</base><ctag>subst:pl:voc:f</ctag></lex>
</tok>
<tok id="tA12">
<orth>Częstochowa</orth>
<lex disamb="1"><base>częstochowa</base>
                <ctag>subst:sg:nom:f</ctag></lex>
</tok>
```

This is a PP syntactically headed by the preposition *dla* with the named entity NP headed by *Huty* 'steel-mill' modified by the proper name *Częstochowa*. Accordingly, there are two constructions here: the NP headed both syntactically and semantically by *Huty*, and the PP, syntactically headed by *dla* and semantically headed by the semantic head of the NP, i.e., by *Huty*.

For the partial annotation stage of the treebank building, we propose a simple standoff annotation consisting of sequence of `<group>` elements containing the information of the extent of the construction (the attributes `from` and `to`), of the syntactic and semantic head (`synh` and `semh`) and of the type of the construction (`PG` for prepositional group and `NG` for nominal group):

```
<group from="tA10" to="tA12" synh="tA10" semh="tA11" type="PG"/>
<group from="tA11" to="tA12" synh="tA11" semh="tA11" type="NG"/>
```

Note that both the syntactic head and the semantic head are tokens (segments) rather than constructions. Since, for each (non-coordinate) construction, the syntactic head is a lexical item, this phrase structure representation can actually be easily translated into dependency representation, in the spirit of [18]. Moreover, instead of saying that the semantic head is the NP argument of the preposition, we are saying that the semantic head of the PP is the semantic head of the NP argument of the PP. This way each construction can be almost (see below) exhaustively characterised by two lexical items within that construction.[10]

The XML representation is more complicated in case of coordination phrases. Such constructions will be marked as `group`s of `type="Coordination"`, without the `synh` and `semh` attributes, but containing `group`s of `type="Conjunction"`, as well as `group`s of `type="Conjunct"`, representing particular conjuncts. For example, assuming that the phrase *zarówno Ratyzbona, jak i Tybinga* 'both Regensburg and Tübingen' is tokenised into 6 segments (*zarówno*, *Ratyzbona*, ,, *jak*, *i*, *Tybinga*) with `id` values from `t1` to `t6`, the partial syntactic annotation may look as follows:

[10] Note that, while we assume that the syntactic head is an immediate constituent of the construction, the semantic head can be deeply embedded, as in the constructed example *[dla [pięciu [największych [z [tych hut]]]]]*, 'for five biggest of these steel-mills', semantically headed by *hut*.

```
<group from="t1" to="t6" type="Coordination"/>
<group from="t1" to="t1" synh="t1" semh="t1" type="Conjunction"/>
<group from="t2" to="t2" synh="t2" semh="t2" type="Conjunct"/>
<group from="t3" to="t5" synh="t5" semh="t5" type="Conjunction"/>
<group from="t6" to="t6" synh="t6" semh="t6" type="Conjunct"/>
```

All (headed; see below) conjuncts provide heads for the whole coordinate structure. Each `group` of `type="Conjunct"` may consist either of a single token (as in the example above), in which case the values of the attributes `from`, `to`, `synh` and `semh` are equal to the `id` of that token, or it may consist of a `group` (simple or coordinate), in which case the values of these attributes are the same as the values of that group. This in particular means that, when one of the conjuncts is a coordinate structure itself, this conjunct will have no `synh` and `semh` attributes, as in the following representation corresponding to the English *either A and B, or C*. Assuming that this construction is tokenised into 7 segments, the representation of such an embedded coordination will be as follows:

```
<group from="t1" to="t7" type="Coordination"/>
<group from="t1" to="t1" synh="t1" semh="t1" type="Conjunction"/>
<group from="t2" to="t4" type="Conjunct"/>
<group from="t2" to="t4" type="Coordination"/>
<group from="t2" to="t2" synh="t2" semh="t2" type="Conjunct"/>
<group from="t3" to="t3" synh="t3" semh="t3" type="Conjunction"/>
<group from="t4" to="t4" synh="t4" semh="t4" type="Conjunct"/>
<group from="t5" to="t6" synh="t6" semh="t6" type="Conjunction"/>
<group from="t7" to="t7" synh="t7" semh="t7" type="Conjunct"/>
```

Any immediate constituent of a coordinate phrase which is neither of the two `type`s above (`Conjunction` or `Conjunct`) is assumed to be a parenthetical, i.e., not the actual part of the coordinate construction.

5 Extending the Poliqarp Query Language

Poliqarp provides a rich query language with two levels of regular expressions: over strings and over segment specifications,[11] but it currently does not make it possible to query a corpus for syntactic representation. It is *not* our aim to extend Poliqarp to a full fledged syntactic query tool; such tools exist, notably the tools created within the TIGER project ([15]; `http://www.ims.uni-stuttgart.de/projekte/TIGER/`). In fact, we have created an XSLT stylesheet converting syntactic information in the format given above (but ignoring the semantic head information) into the TIGER XML format.

However, such general treebank search tools have various restrictions, and the Poliqarp extension described here aims at complementing these tools. One particular restriction of the TIGER tools that the representation described above

[11] For example, the query '`[orth="a{2,}.*[bB]"]{3,}`' could be used to search for sequences of at least three segments whose orthographic form starts with at least two `a`s and ends with a small or capital `b`.

violates is that each node may only have one incoming edge.[12] While the representation above assumes (although it does not enforce) that any given token may be a syntactic head of at most one construction, many constructions may share the same semantic head, as in the example cited in fn. 10 above.

5.1 Simple Constructions

Each segment specification in the Poliqarp query language is a brackets-enclosed combination of constraints connected by logical connectives; for example the following specifies a nominal or adjectival segment whose gender is not feminine:[13] '`[(pos=noun | pos=adj) & gend!=f]`'. Each constraint is an attribute-value specification, where the attribute is either `pos` (part of speech), a grammatical category (e.g., `gend` or `case`), `orth` (orthography) or `base` (the lemma).

Queries for syntactic constructions have a similar syntax, but they use a different repertoire of attributes, non-overlapping with the attributes used to specify segments. Two main attributes to be used for querying for syntactic groups are: `type` and `head`. The attribute `type` refers to the values of the XML attribute `type`, so '`[type=Coordination]`' will find coordinated constructions, while '`[type="[PN]G"]`' will find prepositional and nominal groups.

The syntax of values of the attribute `head` differs from that of the other attributes; its values must be enclosed in a double or a single set of square brackets, as in: '`[head=[...][...]]`' or '`[head=[...]]`'. In the first case, the first brackets specify the syntactic head and second brackets specify the semantic head, as in the following query which may be used to find elective constructions: '`[head=[pos=adj][pos=noun]]`'.

In the second case, the content of the single brackets specifies both the syntactic head and the semantic head and, additionally, makes the requirement that they be the same segment. This means that the queries '`[head=[case=gen][case=gen]]`' and '`[head=[case=gen]]`' have a slightly different semantics: the first will find syntactic groups where the two heads may be different or the same, but they must be genitive; the second will find groups with the two heads being necessarily the same genitive segment.

The usefulness of such queries may be illustrated with a query for verbs which co-occur with dative dependents denoting students; the first approximation of such a query may look like this: '`[pos=verb][head=[case=dat][base=student]]`'. This query will find not only dative nominal groups headed by a form of STUDENT, but also dative numeral groups whose main noun is a form of STUDENT, appropriate dative adjectival elective groups, etc.

Two additional attributes are introduced as syntactic sugar: `synh` and `semh`. The specification '`synh=[...]`' is fully equivalent to '`head=[...][]`', i.e., it puts a constraint on the syntactic head only, while the specification '`semh=[...]`' is fully equivalent to '`head=[][...]`', i.e., no constraint on the syntactic head is given.

[12] There is a special mechanism for adding a second edge, e.g., in order to represent control.

[13] A shorter equivalent query is: '`[pos="noun|adj" & gend!=f]`'.

It may seem that, given the possibility to specify the syntactic head of the construction, the attribute `type` is redundant; in fact, we are not currently aware of cases where the specification '`type="PG"`' or '`type="NG"`' could not be replaced by an appropriate reference to the grammatical class (part of speech) of the syntactic head. However, the `type` attribute is useful for finding constructions which are not defined by their heads, for example, *oratio recta* constructions, and — as we will see below — it is also useful for dealing with coordinate structures.

5.2 Coordination

In §3 we presented the view that coordinate structures are best treated as multi-headed, with each conjunct coming with its own set of syntactic/semantic heads. Given that constructions may have multiple syntactic/semantic head pairs, we give the existential import to specifications like '`[head=[...][...]]`', '`[head=[...]]`', '`[synh=[...]]`' and '`[semh=[...]]`'. That is, a query like '`[head=[pos=noun]]`' will find nominal groups, as well as coordinate groups containing at least one nominal conjunct. The query can be constrained to simple nominal groups or to coordinate constructions by adding an appropriate `type` specification, e.g., '`[head=[pos=noun] & type="NG"]`' should only find simple nominal groups.

This existential semantics of head specifications can be taken advantage of in finding coordinations of unlike categories, as in the query '`[synh=[case=gen] & synh=[case=acc]]`', which may find coordinate phrases with a genitive and an accusative conjunct.[14]

On the other hand, the drawback of this query semantics is that it does not make it possible to find fully homogeneous coordinate structures, with the exclusion of heterogeneous structures mentioned above; i.e., there is currently no way to say that *all* syntactic/semantic head pairs should satisfy a certain requirement. However, the analogy between segment specifications and syntactic group specifications suggests an immediate solution to this problem, namely, allowing an additional operator '`==`' for head specifications, which enforces the universal treatment of the specification. So, just like the query '`[case==gen]`' can be used to search for segments whose all disambiguated interpretations are genitive (cf. §1.3), '`[synh==[pos=noun] & type="Coordination"]`' will find coordinate phrases whose all conjuncts are syntactically nominal groups.

Note that it is theoretically possible that some conjuncts do not have immediate heads; one such situation is illustrated in §4 (p. 56), where the conjunct which is an immediately embedded coordinate structure does not have the attributes `synh` and `semh`. Another such situation may theoretically arise when one of the conjuncts is an *oratio recta* group. In such cases, even if all the other, headed, conjuncts are nominal, the whole coordinate construction will not be identified by the query '`[synh==[pos=noun] & type="Coordination"]`'. However, with the use of the negation operator '`!`', it is possible to formulate a query

[14] Such mixed coordination is possible in Polish in cases where the genitive is actually a partitive genitive realisation of an accusative requirement.

that will find coordinate constructions whose all *headed* conjuncts are nominal, e.g.: '`[synh!=[pos!=noun] & type="Coordination"]`'. This query translates into: find a construction of `type="Coordination"` such that no conjunct can be characterised as having a non-nominal syntactic head; this targets exactly syntactically nominal and headless conjuncts.

6 Conclusion

Although there exist treebanks which contain interesting semantic information, the tectogrammatical level of Prague Dependency Treebank [4] being a good example, to the best of our knowledge few treebanks contain the explicit distinction between syntactic and semantic heads, the Sinica Treebank [10] being the only exception we are aware of. However, both heads must be identified in the process of automatic valence acquisition, as well as in other applications.[15]

This paper gave some rationale for the explicit encoding of such a distinction in a partial treebank of Polish and showed how to implement this encoding: we described how to conservatively extend the XCES encoding to syntactic groups marked with both kinds of heads, and how to conservatively extend the syntax query of Poliqarp to take advantage of this information. Moreover, we proposed a treatment of coordination as multi-headed constructions, and proposed further corresponding extensions of the XML scheme and the Poliqarp query syntax.

The proposal outlined above contains some controversial features, e.g., the identification of heads as segments, i.e., always leaves in the syntactic tree, and the specific treatment of coordination with each conjunct (with the exception of headless conjuncts) bringing its own set of syntactic/semantic heads. However, we feel that ideas presented here are ripe for the community review.

References

1. Florbela Barreto, António Branco, Eduardo Ferreira, Amália Mendes, Maria Fernanda Nascimento, Filipe Nunes, and João Silva. Open resources and tools for the shallow processing of Portuguese: The TagShare project. In *Proceedings of LREC 2006*, 2006.
2. John Beavers and Ivan A. Sag. Coordinate ellipsis and apparent non-constitutent coordination. In Stefan Müller, ed., *Proceedings of the HPSG04 Conference*, pages 48–69, Stanford, CA, 2004. CSLI Publications.
3. Leonard Bloomfield. *Language*. Holt, New York, 1933.
4. Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. The Prague Dependency Treebank: Three-level annotation scenario. In Anne Abeillé, ed., *Treebanks: Building and Using Parsed Corpora*, pages 103–127. Kluwer, Dordrecht, 2003.
5. Oli Christ. A modular and flexible architecture for an integrated corpus query system. In *COMPLEX'94*, Budapest, 1994.

[15] For example, in text retrieval, e.g., [17], in the identification of grammatical relations, e.g., [34], etc.

6. Michael A. Covington. A 700-year-old argument for a syntactic transformation. `http://www.ai.uga.edu/mc/trans700.html`.
7. Jakub Fast and Adam Przepiórkowski. Automatic extraction of Polish verb subcategorization: An evaluation of common statistics. In Zygmunt Vetulani, ed., *Proceedings of the* 2nd Language & Technology Conference, pages 191–195, Poznań, Poland, 2005.
8. Charles J. Fillmore, Collin F. Baker, and Hiroaki Sato. Seeing arguments through transparent structures. In *Proceedings of LREC 2002*, pages 787–791, Las Palmas, Canary Islands, Spain, 2002. ELRA.
9. Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck. Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250, 2003.
10. Chu-Ren Huang, Chen Keh-Jiann, Chen Feng-Yi, Chen Keh-Jiann, Gao Zhao-Ming, and Chen Kuang-Yu. Sinica treebank: Design criteria, annotation guidelines, and on-line interface. In *Proceedings of 2nd Chinese Language Processing Workshop (Held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics, ACL-2000)*, pages 29–37, Hong Kong, 2000.
11. Nancy Ide, Patrice Bonhomme, and Laurent Romary. XCES: An XML-based standard for linguistic corpora. In *Proceedings of the Linguistic Resources and Evaluation Conference*, pages 825–830, Athens, Greece, 2000.
12. Daniel Janus and Adam Przepiórkowski. Poliqarp 1.0: Some technical aspects of a linguistic search engine for large corpora. In Jacek Waliński, Krzysztof Kredens, and Stanisław Goźdź-Roszkowski, eds., *The proceedings of Practical Applications of Linguistic Corpora 2005*, Frankfurt am Main, 2006. Peter Lang. To appear.
13. Krystyna Kallas. *Składnia współczesnych polskich konstrukcji współrzędnych*. Wydawnictwo Uniwersytetu Mikołaja Kopernika, Toruń, 1993.
14. Iwona Kosek. *Przyczasownikowe frazy przyimkowo-nominalne w zdaniach współczesnego języka polskiego*. Wydawnictwo Uniwersytetu Warmińsko-Mazurskiego, Olsztyn, 1999.
15. Wolfgang Lezius. TIGERSearch — ein Suchwerkzeug für Baumbanken. In Stephan Busemann, ed., *Proceedings der 6. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2002)*, Saarbrücken, 2002.
16. Igor A. Mel'čuk. Levels of dependency in linguistic description: concepts and problems. In Vilmos Àgel, Ludwig Eichinger, Hans-Werner Eroms, Peter Hellwig, Hans-Jürgen Heringer, and Henning Lobin, eds., *Dependenz und Valenz: Ein Internationales Handbuch Der Zeitgenösischen Forschung*, pages 188–229. De Gruyter, Berlin, 2003.
17. Christof Monz and Maarten de Rijke. Tequesta: The University of Amsterdam's texual question answering system. In *Proceedings of Tenth Text Retrieval Conference (TREC-10)*, pages 513–522, 2001.
18. Joakim Nivre. Theory-supporting treebanks. In Joakim Nivre and Erhard Hinrichs, eds., *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT2003)*, pages 117–128, Växjö, Norway, 2003.
19. Carl Pollard and Ivan A. Sag. *Information-Based Syntax and Semantics, Volume 1: Fundamentals*. Number 13 in CSLI Lecture Notes. CSLI Publications, Stanford, CA, 1987.
20. Carl Pollard and Ivan A. Sag. *Head-driven Phrase Structure Grammar*. Chicago University Press / CSLI Publications, Chicago, IL, 1994.
21. Adam Przepiórkowski. *Case Assignment and the Complement-Adjunct Dichotomy: A Non-Configurational Constraint-Based Approach*. Ph. D. dissertation, Universität Tübingen, 1999.

22. Adam Przepiórkowski. *The IPI PAN Corpus: Preliminary version*. Institute of Computer Science, Polish Academy of Sciences, Warsaw, 2004.
23. Adam Przepiórkowski. On heads and coordination in a partial treebank. In Jan Hajič and Joakim Nivre, eds., *Proceedings of the TLT 2006*, pages 163–174, Prague, 2006.
24. Adam Przepiórkowski and Jakub Fast. Baseline experiments in the extraction of Polish valence frames. In Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, eds., *Intelligent Information Processing and Web Mining*, Advances in Soft Computing, pages 511–520. Springer-Verlag, Berlin, 2005.
25. Adam Przepiórkowski, Zygmunt Krynicki, Łukasz Dębowski, Marcin Woliński, Daniel Janus, and Piotr Bański. A search tool for corpora with positional tagsets and ambiguities. In *Proceedings of LREC 2004*, pages 1235–1238, Lisbon, 2004. ELRA.
26. Adam Przepiórkowski and Marcin Woliński. A flexemic tagset for Polish. In *Proceedings of* Morphological Processing of Slavic Languages*, EACL 2003*, pages 33–40, Budapest, 2003.
27. Adam Przepiórkowski and Marcin Woliński. The unbearable lightness of tagging: A case study in morphosyntactic tagging of Polish. In *Proceedings of the* LINC-03*, EACL 2003*, pages 109–116, 2003.
28. Ivan A. Sag, Gerald Gazdar, Thomas Wasow, and Steven Weisler. Coordination and how to distinguish categories. *Natural Language and Linguistic Theory*, 3:117–171, 1985.
29. Zygmunt Saloni and Marek Świdziński. *Składnia współczesnego języka polskiego*. Wydawnictwo Naukowe PWN, Warsaw, 4th (changed) edition, 1998.
30. Petr Sgall, Eva Hajičová, and Jarmila Panevová. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Reidel, Dordrecht, 1986.
31. Max Silberztein. Finite-state description of the French determiner system. *French Language Studies*, 13:221–246, 2003.
32. Marek Świdziński. Realizacje zdaniowe podmiotu-mianownika, czyli o strukturalnych ograniczeniach selekcyjnych. In Andrzej Markowski, ed., *Opisać słowa*, pages 188–201. Dom Wydawniczy Elipsa, 1992.
33. Lucien Tesnière. *Éléments de Syntaxe Structurale*. Klincksieck, Paris, 1959.
34. Rebecca Watson, John Carroll, and Ted Briscoe. Efficient extraction of grammatical relations. In *Proceedings of the Ninth International Workshop on Parsing Technology*, pages 160–170, Vancouver, British Columbia, 2005. Association for Computational Linguistics.
35. Abby Wright and Andreas Kathol. When a head is not a head: A constructional approach to exocentricity in English. In Jonh-Bok Kim and Stephen Wechsler, eds., *Proceedings of the 9th International Conference on Head-Driven Phrase Structure Grammar*, pages 373–389. CSLI Publications, Stanford, CA, 2003.