Syntactic processing of the IPI PAN Corpus of Polish

Adam Przepiórkowski, Aleksander Buczyński, Daniel Janus (Institute of Computer Science, Polish Academy of Sciences)

The aim of this paper is to present recent and ongoing work on adorning the IPI PAN Corpus of Polish (Przepiórkowski 2004, 2006a) with partial syntactic annotation, with the ultimate aim of building a treebank of Polish. The work described here is a part of the project *Automatic extraction of linguistic knowledge from a large corpus of Polish* (a Ministry of Education and Science grant number 3T11C00328), aiming at the automatic construction of a valence dictionary.

Treebanks are often built manually or semi-manually, in the excruciating process, where highly qualified linguists painstakingly add full syntactic analyses to corpus sentences. In the case of the IPI PAN Corpus, we decided to divide the task into a sequence of simpler tasks, where the first task consists in the automatic identification of various types of simple phrases, without any attempt at solving attachment ambiguities: this task is known as surface, shallow or partial parsing. On the other hand, as explained in detail in Przepiórkowski 2006b, 2007a, automatic valence acquisition requires that the phrases be marked not only with their syntactic heads, but also with semantic heads; in case of most of noun phrases, the syntactic head noun is also the semantic head of the phrase, but, e.g., in Polish numeral phrases, syntactic heads are numerals, while semantic heads are expressed by sisters (complements) of these numerals.

To the best of our knowledge, there are no publicly available partial parsers of Polish, while the only publicly available deep parser of Polish, Świgra (Woliński 2004), has not been evaluated on naturally occurring texts and it does not provide information about semantic heads. For these reasons, we decided to develop a partial grammar of Polish, suited to the task at hand. While the grammar is currently under development, a stable version will be ready for presentation by the time of PLM 2007. The grammar is perhaps unique among shallow grammars in that it does not assume a morphosyntactically disambiguated input. On the contrary, the formalism for developing the grammar (devised by the first author and implemented by the second author; cf. Przepiórkowski 2007b and Buczyński 2007) allows for simultaneous parsing and morphosyntactic disambiguation. A trivial example of one rule of such a grammar is given below:

```
Left: [pos~~"prep"]
Match: [pos~~"num"]
       [pos~~"adj"]*
       [pos~~"subst"]
Actions: unify(case,1,2,3,4),
         unify(number gender,2,3,4),
         group(NumG,2,4)
```

This rule finds, in the right context of a preposition, a sequence of a numeral, followed by any number of adjectives and a noun, tries to unify the case of all the segments in the match and the preceding preposition (the case of a preposition is the case it assigns to its argument), unify the number and gender of all the segments in the match, and if these unifications succeed, it marks the match as a numeral group, with the numeral as its syntactic head, and the noun as its semantic head. A side effect of this rule is the rejection of those morphosyntactic interpretations which do not satisfy the unify conditions.

The IPI PAN Corpus is annotated with a linguistically motivated and worked out tagset (Przepiórkowski and Woliński 2003), and the corpus search tool developed for the IPI PAN Corpus, Poliqarp (Przepiórkowski et al. 2004, Janus and Przepiórkowski 2006, Janus 2006) implements a very rich query language over that tagset, but it was not developed with the aim of searching a syntactically annotated corpus. Such general treebank searching tools exist, e.g., http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/, but they are constrained in ways which are incompatible with the distinction between syntactic and semantic heads. For this reason, a syntactic extension to the Poliqarp query language has been designed (by the first author) and implemented (by the third author; cf. Janus and Przepiórkowski 2007), and we plan to present it at the *Advances in natural language and speech processing* workshop at PLM 2007.

References:

Aleksander Buczyński. (2007). An effective implementation of combined partial parser and morphosyntactic disambiguator. ACL 2007 Student Research Workshop, Prague.

Daniel Janus. (2006). Metody przeszukiwania dużych korpusów tekstów. M.Sc. Thesis, Faculty of Mathematics, Informatics and Mechanics, Warsaw University, Warsaw. [http://korpus.pl/~nathell/praca.pdf]

Daniel Janus and Adam Przepiórkowski. (2006). POLIQARP 1.0: Some technical aspects of a linguistic search engine for large corpora. In the proceedings of PALC 2005, Peter Lang, Frankfurt am Main. [http://nlp.ipipan.waw.pl/~adamp/Papers/2006-poliqarp/]

Daniel Janus and Adam Przepiórkowski. (2007). Poliqarp: An open source corpus indexer and search engine with syntactic extensions. ACL 2007 Demo Session.

Adam Przepiórkowski. (2004). Korpus IPI PAN. Wersja wstępna / The IPI PAN Corpus: Preliminary version. Institute of Computer Science, Polish Academy of Sciences, Warsaw. [http://nlp.ipipan.waw.pl/~adamp/Papers/2004-corpus/]

Adam Przepiórkowski. (2006a). The Potential of the IPI PAN Corpus. In: PSiCL 41, pp. 31-48. [http://nlp.ipipan.waw.pl/~adamp/Papers/2005-psicl-numbers/]

Adam Przepiórkowski. (2006b). What to acquire from corpora in automatic valence acquisition. In: Violetta Koseska-Toszewa and Roman Roszko, eds., *Semantyka a konfrontacja językowa*, 3. [http://nlp.ipipan.waw.pl/~adamp/Papers/2006-what.to.acquire/what.to.acquire.pdf]

Adam Przepiórkowski. (2007a). On Heads and Coordination in Valence Acquisition. In: Alexander Gelbukh, ed., *Computational Linguistics and Intelligent Text Processing* (CICLing 2007), Springer Verlag, LNCS series, pp. 50-61. [http://nlp.ipipan.waw.pl/~adamp/Papers/2007-cicling/43940050.pdf]

Adam Przepiórkowski. (2007b). A Preliminary Formalism for Simultaneous Rule-Based Tagging and Partial Parsing. In: Georg Rehm, Andreas Witt and Lothar Lemnitzer, eds., *Data Structures for Linguistic Resources and Applications: Proceedings of the Biennial GLDV Conference 2007*, Gunter Narr Verlag, pp. 81-90. [http://nlp.ipipan.waw.pl/~adamp/Papers/2007-gldv/final.pdf]

Adam Przepiórkowski, Zygmunt Krynicki, Łukasz Dębowski, Marcin Woliński, Daniel Janus and Piotr Bański. (2004). A Search Tool for Corpora with Positional Tagsets and Ambiguities. In the Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, pp. 1235-1238. [http://nlp.ipipan.waw.pl/~adamp/Papers/2004-lrec/]

Adam Przepiórkowski and Marcin Woliński. (2003). The Unbearable Lightness of Tagging: A Case Study in Morphosyntactic Tagging of Polish. In: The Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03), EACL 2003. [http://nlp.ipipan.waw.pl/~adamp/Papers/2003-eacl-ws03/]

Marcin Woliński. (2004). Komputerowa weryfikacja gramatyki Świdzińskiego. Ph.D. Thesis, Institute of Computer Science, Polish Academy of Sciences, Warsaw.