Abs-192

Adam Przepiórkowski (Institute of Computer Science, PAS), Rafał L. Górski (Institute of Polish Language, PAS), Marek Łaziński (University of Warsaw), and Piotr Pęzik (University of Łódź)

The National Corpus of Polish: benefits of synergy

In this paper we report more than three years of work on the National Corpus of Polish. What makes the National Corpus of Polish project different from a typical YACP (Yet Another Corpus Project)? Before the project started there were four corpora of Polish available. Each had some merits but none of them met all the requirements (large, balanced, annotated, publicly available) of a modern general reference corpus.

The NCP is an effect of a joint effort of four teams which constructed the mentioned corpora: the Institute of Computer Science PAS, Institute of Polish Language PAS, Chair of English (University of Lodz) and Polish Scientific Publishers PWN. Thus – contrary to most national corpora – the work did not start from scratch. Each of the teams brought expertise, but also their resources and tools. The project however went far beyond simply merging the four corpora.

The result of the project is a large corpus of over 1.5 billion tokens, a 300 million balanced subcorpus, and a 1 million manually annotated subcorpus.

In the course of the project a number of tools have been developed, including Anotatornia (an on-line tool for manual annotation of texts), two search engines, a morphosyntactic tagger, tools for word sense disambiguation, named entity recognition and shallow syntactic parsing. Last but not least: a ready solution of XML annotation (based on TEI P5) which is well documented. All these tools are publicly available and can be (or in case of Poliqarp are already) reused.

Still the corpus was compiled not only with a view to Natural Language Processing, but also to purely linguistic applications. Every year a demo version of the corpus was launched, allowing for feedback from the users. This feedback was quite intensive because the corpus, also at the stage of compiling, was an empirical basis of a large dictionary of modern Polish. The corpus was also used for educational purposes, including training translators.

Due to the considerable size of its unbalanced part, the corpus is also treated as a repository of texts which can be rearranged to form a new corpus. Two examples are: a corpus of Polish of the 21st century as a counterpart of a corpus of the sixties for tracking short-term diachronic processes, and a corpus of Polish directly comparable to BNC.

The project has proven that merging various corpora compiled for one language is worth the effort. The effect is much more than simply a sum of the scattered corpora, but provides a corpus of an entirely new quality.

References:

Acedański, S. (2010). A morphosyntactic brill tagger with lexical rules for inflectional languages. In Advances in Natural Language Processing: Proceedings of the 7th International Conference on Natural Language Processing, IceTAL 2010, Reykjavík, Iceland, Lecture Notes in Artificial Intelligence, Berlin. Springer-Verlag.

Górski, R.L. (2008): Representativeness of a written part of a Polish general-reference corpus. Primary notes. [w:] Barbara Lewandowska-Tomaszczyk (red.): Corpus Linguistics, Computer Tools, and Applications - State of the Art. PALC 2007, Frankfurt/M etc: Peter Lang.

Przepiórkowski, A. and Murzynowski, G. (to appear). Manual annotation of the National Corpus of Polish with Anotatornia.

Pęzik, P. (to appear). Providing corpus feedback for translators with the PELCRA search engine for NKJP.

Savary, A., Waszczuk, J., and Przepiórkowski, A. (2010). Towards the annotation of named entities in the National Corpus of Polish. In Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010, Valletta, Malta. ELRA.