New linguistic annotation in the National Corpus of Polish

It has been eight years since the official release of the National Corpus of Polish (NKJP, Przepiórkowski et al., 2012). Despite efforts aimed at raising new funds and rebooting the project, the corpus remained unchanged since 2012 and the most recent texts included in the resource date back to 2010. Yet NKJP remains the most popular general purpose corpus of Polish. The aim of this talk is to present the existing balanced and representative 300 mln tokens large corpus in a refreshed form much similar to the experience known to some users from Korpusomat web application (Kieraś et al., 2018).

First of all, the corpus was morphologically analysed and tagged anew using current versions of Morfeusz SGJP analyser (Woliński, 2014) and Conraft-pl tagger (Waszczuk 2012; Waszczuk et al., 2018). The analyser's dictionary provides an extensive lexical coverage of the data and the tagger brings higher disambiguation accuracy (approx. 94%), which makes the corpus queries more reliable comparing to the 2012 standards. Some changes in the morphosyntactic tagset were also made, especially regarding the grammatical gender. The changes were aimed at bringing some distinctions closer to those known from the Grammatical Dictionary of Polish (Saloni et al., 2015).

Apart from that, the corpus annotation was extended with two new layers: a named entities layer and a syntactic layer. The former is based on the named entities description from the original NKJP project and performed automatically on the full corpus using Liner2 classifier (Marcińczuk et al., 2017). The latter is based on the results of dependency parsing performed by Combo parser (Rybak and Wróblewska, 2018) and based on the Polish Dependency Bank annotation scheme. Since it was not possible to index full dependency trees in the corpus search engine, we have decided to provide the user with only partial syntactic annotation — for each token in the corpus the syntactic layer provides information about its syntactic head (immediate predecessor) in the dependency tree (the head's lemma and inflectional tag as well as the distance and left or right position relative to the queried word).

The multi-layer corpus annotation was indexed with MTAS web-based search engine (Brouwer et al., 2016) which allows for querying the corpus using Corpus Query Language (CQL) known from numerous other corpus engines. The web interface provides a simple graphical interface for query construction. It is possible to refer to any number of layers simultaneously in one query. Some possible applications of such queries will be shown in the presentation.

The NKJP version described in this abstract is currently under development and will be publicly available on the Web later this year.

References

Brouwer, M., Brugman, H., Kemps-Snijders, M, MTAS: A Solr/Lucene based Multi Tier Annotation Search solution, Selected papers from the CLARIN Annual Conference 2016.

Kieraś, W., Kobyliński, Ł., Ogrodniczuk, M., Korpusomat — a tool for creating searchable morphosyntactically tagged corpora. Computational Methods in Science and Technology, 24(1):21–27, 2018.

Marcińczuk, M., Kocoń, J., Oleksy, M., Liner2 - a Generic Framework for Named Entity Recognition, In Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, pages 86-91, Valencia, Spain, 2017.

Przepiórkowski, A., Bańko, M., Górski, R. L., Lewandowska-Tomaszczyk, B., Narodowy Korpus Języka Polskiego, Warszawa 2012.

Rybak, P., Wróblewska, A., Semi-supervised neural system for tagging, parsing and lematization. In Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 45–54. Association for Computational Linguistics, 2018.

Saloni, Z., Woliński, M., Wołosz, R., Gruszczyński, W., Skowrońska, D. Słownik gramatyczny języka polskiego. Warszawa, 3rd edition, 2015.

Waszczuk, J., Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language. In Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012), pages 2789–2804, Mumbai, India, 2012.

Waszczuk, J., Kieraś, W., Woliński, M., Morphosyntactic disambiguation and segmentation for historical Polish with graph-based conditional random fields. In: Petr Sojka, Aleš Horák, Ivan Kopeček, Karel Pala, editors, Text, Speech, and Dialogue: 21st International Conference, TSD 2018, Brno, Czech Republic, September 11-14, 2018, Proceedings, number 11107 in Lecture Notes in Artificial Intelligence, pages 188–196. Springer-Verlag, 2018.

Woliński, M., Morfeusz reloaded. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, pages 1106–1111, Reykjavík, Iceland, 2014. ELRA.