Korpusomat (‘corpus machine’, korpusomat.pl) is a web application aimed at building automatically indexed and annotated searchable corpora from documents provided by the user. Korpusomat integrates various natural language processing tools and provides an intuitive interface to use them in own projects.

Since its very first version in 2016 Korpusomat has gained a positive response from the Polish corpus linguistics community and was addressed with many feature and enhancement requests. The application is thus in constant development. It has been recently equipped with new search engine, enabling searching various layers of annotation (Brouwer et al., 2017). Each text sent to Korpusomat is automatically annotated with Morfeusz morphological analyzer (Woliński, 2014), with one of the two alternative dictionaries: SGJP (Saloni et al., 2015) or Polimorf (Woliński et al., 2012), and one of the two morphosyntactic disambiguating taggers: Concraft (Waszczuk, 2012) or Toygger (Krasnowska-Kieraś, 2017), each representing different technical approach to the problem. The MTAS search engine allows to query the annotated corpus using Corpus Query Language (CQL) which is familiar to Sketch Engine and National Corpus of Polish users. Query results may be downloaded in CSV format for further off-line processing, i.e. using advanced statistical tools such as R or simply in Excel spreadsheets.

Korpusomat accepts files in many formats: from plain text files and Word DOC(X) files to e-book EPUB and MOBI formats and two layer DJVU files. Each document can be described by metadata, both predefined and user defined. Some metadata fields are automatically completed if the information was provided by the source format. The metadata entries can be later used in searching and providing basic statistics concerning the corpus.

The presentation will also give some brief overview of development plans. Future development will focus mainly on integrating other layers of automatic annotation and data extraction, such as named entities, dependency parsing, semantic roles etc. Thus it strongly relies on the state of development of such tools for Polish and their accuracy. It is our intention to make Korpusomat a standard, easy to deploy framework for all kinds of static corpora of Polish texts, such as parliamentary corpus, historical corpora and all sorts of special purpose corpora collected as a basis for various linguistic research and projects. It would make any technical updates easier and would allow keeping previously collected corpora processed with the most up-to-date NLP tools.

References

Brouwer, M., Brugman, H., and Kemps-Snijders, M. (2017). MTAS: A Solr/Lucene based Multi Tier Annotation Search solution. In Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 26–28 October 2016, CLARIN Common Language Resources and Technology Infrastructure, number 136, pages 19–37. Linköping University Electronic Press, Linköpings universitet.

Krasnowska-Kieraś, K. (2017). Morphosyntactic disambiguation for Polish with bi-LSTM neural networks. In Proceedings of 8th Language & Technology Conference, pages 367–371.

Marciniak, M., Mykowiecka, A., and Rychlik, P. (2016). TermoPL — a flexible tool for terminology extraction. In Calzolari, N., Choukri, K., Declerck, T., Grobelnik, M., Maegaard, B., Mariani, J., Moreno, A.,

Odijk, J., and Piperidis, S., editors, Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016, pages 2278–2284, Portorož, Slovenia. ELRA, European Language Resources Association (ELRA).

Saloni, Z., Gruszczyński, W., Woliński, M., Wołosz, R., and Skowrońska, D. (2015). Słownik gramatyczny języka polskiego. 3. edition.

Waszczuk, J. (2012). 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), Mumbai, India.

Woliński, M. (2014). Morfeusz reloaded. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, pages 1106–1111, Reykjavík, Iceland. ELRA.

Woliński, M., Miłkowski, M., Ogrodniczuk, M., Przepiórkowski, A., and Szałkiewicz, Ł. (2012). PoliMorf: a (not so) new open morphological dictionary for Polish. In Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, pages 860–864, Istanbul, Turkey. ELRA.