Automatic generation of the Estonian Collocation Dictionary database

Jelena Kallas, Adam Kilgarriff, Kristina Koppel, Elgar Kudritski, Jan Michelfeit, Maria Tuulik, Ülle Viks

Keywords: corpus lexicography, automatic generation, collocation dictionary, corpus query system, dictionary writing system, Estonian language

The paper reports on the process of the automatic generation of the Estonian Collocation Dictionary (ECD) database. The database has been compiled by the Institute of the Estonian Language in collaboration with Lexicon Computing Ltd.

ECD is a monolingual online scholarly dictionary aimed at learners of Estonian as a foreign or second language at the upper intermediate and advanced levels. The dictionary contains about 10,000 headwords, including single items and multi-word lexical items. The collocates within each headword are grouped according to the lexico-grammatical structure formed by the collocational phrase, and for each collocation one or two example sentences are provided. The ECD project started in 2014 and the dictionary is scheduled to be published in 2018.

For the automatic generation of the ECD database, the corpus query system Sketch Engine (Kilgarriff et al. 2004) Word List, Word Sketch and Good Dictionary Example (GDEX) functions were used. The data were automatically extracted in an XML format from the 463-million-word Estonian National Corpus (https://the.sketchengine.co.uk/auth/corpora/) and imported into the XML-based EELex dictionary writing system (Langemets et al. 2006, Jürviste et al. 2011). To make the importing of automatically extracted data from Sketch Engine into EELex possible, the XML structure for extracted data was matched with the XML structure of ECD in EELex.

We implemented the methodology proposed by Kosem et al. 2013. The procedure required the following: a selection of lemmas, finely-grained Sketch Grammar, GDEX (Kilgarriff et. al 2008) configuration, the API script to extract data from Word Sketch and settings for extraction. The list of lemmas was compiled using the Word List function. The latest Sketch Grammar version 1.6 was developed and improved; it includes all of the lexico-grammatical structures that will be presented in the ECD. Grammar contains 116 rules in total. For the extracting of dictionary examples, the first version of GDEX for Estonian was developed. Classifiers concerning sentence optimum length, word optimum length, number of punctuation marks, word frequency, lemma repetition, anaphors, tokens with capital letters and symbols, abbreviations and a list of "bad words" were proposed and implemented. The use of classifiers brought significant improvements to the output.

For automatic extraction, the following parameters were specified: a list of gramrels, minimum frequency and salience of gramrels, the number of collocates per grammatical relation, the minimum frequency and salience of a collocate, and the number of examples per collocate.

As a result, the database contains 10 939 headwords, 82 678 gramrels, 493 971 collocates and 2 469 855 example sentences (five example sentences for each collocate). Additionally, the

database includes the part-of-speech and overall frequency number of each headword, the overall frequency of each gramrel and collocate, and the score of each gramrel and collocation. Currently, the database is being examined, edited and supplemented by lexicographers.

In the paper we will also discuss the possible implementation of automatically extracted statistical data for the visualisation of collocational information in order to facilitate dictionary navigation options.

References:

Kilgarriff, A., Rychly, P., Smrž, P., Tugwell, D. 2004. The Sketch Engine. In: G. Williams, S. Vessier (eds.). Proceedings of the XI Euralex International Congress. Lorient: Université de Bretagne Sud, pp. 105–116.

Kilgarriff, A., Husák, M., McAdam, K., Rundell M., Rychlý, P. 2008. GDEX: Automatically finding good dictionary examples in a corpus. In: E. Bernal, J. DeCesaris (eds.). Proceedings of the 13th EURALEX International Congress. Barcelona: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra, pp. 425–432.

Kosem, I., Gantar, P., Krek, S. 2013. Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. In: I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, M. Tuulik (Eds.). Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia, pp. 17–19.

Jürviste, M., Kallas, J., Langemets, M., Tuulik, M., Viks, Ü. 2011. Extending the functions of the EELex dictionary writing system using the example of the Basic Estonian Dictionary. In: I. Kozem, K. Kozem (eds.). eLexicography in the 21st Century: New Applications for New Users, Proceedings of eLex 2011, Bled, 10-12 November 2011. Ljubljana: Trojina, Institute for Applied Slovenian Studies, pp. 106–112.

Langemets, M., Loopmann, A., Viks, Ü. 2006. The IEL dictionary management system of Estonian. In: G.-M. de Schryver (ed.). DWS 2006: Proceedings of the Fourth International Workshop on Dictionary Writing Systems: Pre-EURALEX workshop: Fourth International Workshop on Dictionary Writing System. Turin, 5th September 2006. Turin: University of Turin, pp. 11–16.