Estonian Collocations Dictionary: conception and implementation

Jelena Kallas, Maria Tuulik
Institute of the Estonian Language
E-mail: jelena.kallas@eki.ee, maria.tuulik@eki.ee

1 Introduction

The Estonian Collocations Dictionary will be a monolingual online, corpus-driven, scholarly dictionary aimed at learners of Estonian as a foreign language or second language at the upper intermediate and advanced levels (B2 to C1) according to the Common European Framework of Reference for Languages. The dictionary contains about 10,000 headwords, including single items and multi-word lexical items. The dictionary will be compiled in the dictionary writing system EELex (Jürviste et al. 2011).The most frequent collocation patterns will be analysed using the Sketch Engine corpus query system (Kilgarriff et al. 2004).
The project started in 2014 and it should be completed by 2018. However, we plan partial publication of edited entries as part of the online Basic Estonian Dictionary (Kallas et al 2014). The Basic Estonian Dictionary will be published online in September 2014.

2 Lexicographical work flow

The project is planned to last four years.

Phase	**Duration**
Preparation	2014
Data acquisition	2014
Computerisation	2014
Data processing	2015-2017
Data analysis	2015-2017
Preparation for online release	on-going
Afterlife	Linking the resource with other scholarly dictionaries and language-learning environments

Table 1 Process phases of the dictionary project and their time spans

2.1 Preparation

In this phase, we will develop the general conception of the dictionary, prepare the Style Guide (Atkins, Rundell 2008: 118-123), and begin to develop the database structure. The dictionary writing system EELex is XML-based. The most difficult aspect is to decide how to link new data with previously compiled resources.
We still lack a clear strategy of how to monitor users and how to reattach them to the dictionary. There is very low awareness of the electronic resources produced by the Institute of the Estonian Language (mostly language professionals use it). So far we have created a Facebook homepage and we might implement a technique such as "word of the day" to be posted on Facebook every day to keep potential users' attention on our resource.

2.2 Data acquisition

The primary source of the dictionary will be the recently compiled Estonian National Corpus (560 mln tokens). The corpus consists of the Estonian Reference Corpus (contains texts written up to 2008) and the Estonian Web Corpus etTenTen (350 mln tokens). etTenTen was compiled by Lexical Computing Ltd. in 2013. The corpus was annotated morphologically, lemmatized, partially disambiguated and annotated by clauses by Filosoft LLC and installed into Sketch Engine software. We have also analysed all other resources suitable for the investigation of the collocational properties of headwords. Since the dictionary is meant for the development of writing skills, we are particularly interested in academic writing vocabulary lists.

2.3 Computerisation

We plan to use a semi-automatic approach for database compilation. One possibility is the Tickbox Lexicography method (Kilgarriff et al. 2010). In this stage, we plan to set up the TBL in such a way that it will be possible to import relevant collocates and example sentences from Sketch Engine word sketches into EELex.
In order to do this, we need to elaborate Sketch Grammar and investigate classifiers needed for Estonian GDEX (Kilgarriff et al. 2008; see also Kozem et al. 2013). Estonian Sketch Grammar (Kallas 2013) contains 85 rules

(14 UNARY, four SYMMETRIC, 62 DUAL and five TRINARY grammatical relations). As a result, the system searches for 32 types of lexico-grammatical constructions. The dictionary will be designed in the same way. Collocates will be grouped according to the lexico-grammatical structures formed by collocational phrases. This approach eliminates the necessity of analysing grammatical and lexical items separately. It also allows for a more systematic use of word sketches.

2.4 Data processing and data analysis

We will start with data analysis as soon as the phase of computerization is completed. Since we will use Tickbox Lexicography as the compilation method, the stage of lexicographic data processing will be merged with the stage of data analysis. We will use the results of automatic analysis concerning grammatical information, collocations and example sentences, but we will choose manually what kind of information will be given in a particular entry. Separate work needs to be done on the lemma list and multi-word headwords.

2.5 Preparation for online release

After proofreading, we will publish the dictionary. In this phase, a lot of the work can be done automatically. EELex automatically checks if there are broken links. We can also use the bulk correction function if it is necessary to make global changes inside all of the entries.

2.6 Afterlife

In this phase, we plan to link the dictionary with other scholarly dictionaries and language-learning environments.

3 References

Atkins, Sue; Rundell, Michael 2008. Oxford Guide to Practical Lexicography. Oxford: Oxford University Press.

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.

Kallas, J., Tuulik, M., Langemets, M. 2014 (forthcoming). The Basic Estonian Dictionary: the first monolingual L2 learner's dictionary of Estonian. – Proceedings of the XVI Euralex Congress.

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, Adam; Husák, Milos; McAdam, Katy; Rundell Michael; Rychlý, Pavel 2008. GDEX: Automatically finding good dictionary examples in a corpus. – E. Bernal, J. DeCesaris (eds.) Proceedings of the 13th EURALEX International Congress. Barcelona: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra, pp. 425–432.

Kilgarriff, Adam; Kovář, Vojtěch; Rychlý, Pavel 2010. Tickbox lexicography. – S. Granger, M. Paquot (eds.). eLexicography in the 21st century: New challenges, new applications. Proceedings of eLex 2009, Louvain-la-Neuve, 22-24 October 2009. Louvain-la-Neuve: Presses universitaires de Louvain, 411–418.

Kallas, Jelena 2013. Eesti keele sisusõnade süntagmaatilised suhted korpus- ja õppeleksikograafias. [Syntagmatic relationships of Estonian content words in corpus and pedagogical lexicography.] Tallinn: Tallinn University. Dissertations on Humanities Sciences.

Kosem, Iztok; Husák, Milos; McCarthy, Diana 2011. GDEX for Slovene. – Proceedings of eLex 2011, 151–159