# New Estonian Words and Senses: Detection and Description

**Margit Langemets, Jelena Kallas, Kaisa Norak, Indrek Hein**

Institute of the Estonian Language

margit.langemets@eki.ee, jelena.kallas@eki.ee, kaisa.norak@gmail.com, indrek.hein@eki.ee

## Abstract

The Web era has brought about an urgent need for the automatic monitoring of language, including the extraction of new words and senses. In order to monitor language, especially lexical changes, the Institute of the Estonian Language, in cooperation with Lexical Computing Ltd., crawls the Web every two years. Corpora are used through the corpus query system Sketch Engine (Kilgarriff et al. 2004)[1] and CQS KORP[2]. The most recent corpus is the Estonian Reference Corpus 2017 (1.1 billion words); the next corpus will be crawled in 2019. We also implement crowdsourcing techniques for neologism registration by offering our users the opportunity to propose new words or senses. They can do this by using the feedback forms on our dictionary portals Sõnaveeb ('Wordweb')[3] and e-keelenõu ('e-Language advice')[4].

In this paper, we present the results of an experimental study on neologism detection on the basis of text collection, which was compiled at the Institute from 2016 to 2018. We describe the method for neologism detection and evaluate the results. This is the first study for Estonian aimed at the development of a tool to supply lexicographers with neologism candidates for inclusion in a dictionary.

---

[1] https://sketchengine.eu/ (accessed March 30, 2019)

[2] https://korp.keeleressursid.ee/ (accessed March 30, 2019)

[3] https://sonaveeb.ee (accessed March 30, 2019)

[4] http://keeleabi.eki.ee/ (accessed March 30, 2019)

We discuss the practice of providing both prescriptive and descriptive information about new words. The prescriptive data concerns mostly orthography and inflection and should indicate what belongs to standard Estonian and what does not. However, dealing with neologisms is no trivial task. Within the unified single database Ekilex[5], we will present both descriptive and prescriptive data.

**Keywords**

neologisms, corpus lexicography, dictionary portal, Estonian

**Introduction**

There is quite a long tradition of dictionaries of new words in Estonian lexicography, beginning with Grenzstein (1884, 1,600 words) and Aavik (1919, 2nd ed. 1921, 4,000 words). All new printed editions of monolingual dictionaries contain a selection of new words, although collections of new words have also been published separately (e.g. Erelt, Meriste and Kull 1985) alongside large dictionaries, as in the printing era the compilation of large dictionaries took decades to complete.

The situation has changed rapidly since the 2000s, especially with the proliferation of Web 2.0 and the Social Web, which includes wikis, blogs, apps, collaborative platforms etc. This has also affected lexicography to a large extent. Since 2000 we have experienced a transitional period in Estonia, where dictionaries have been published on paper as well as electronically (e.g. EKSS 2009 and ÕS 2013, 2018). These online versions have been almost exact copies of the paper dictionary.

Within the framework of the new dictionary-writing system Ekilex (Tavast et al. 2018), we are moving on from presenting separate interfaces for different dictionaries to unified data in order to provide data in an aggregated form. The long-term vision is to have a

[5] https://ekilex.eki.ee (accessed March 30, 2019)

single data source (Ekilex) that provides (also via the API) consistent and comprehensive information about Estonian words, combining the research carried out in all departments and working groups of the Institute. The system serves to provide data for dictionary users via the new language portal Sõnaveeb (‘Wordweb’), as well as for lexicographers working on different tasks. The new language portal Sõnaveeb (‘Wordweb’) was released in February 2019.

One clear advantage of a single data source is the opportunity to constantly update different data subsets in the general database. So, once a new word has been added to the database by lexicographers dealing with contemporary Estonian, this word can be immediately supplied with a Russian, Finnish, Hungarian or other language translation equivalent provided by lexicographers dealing with bilingual datasets.

In the next sections, we will discuss both the methods used so far and the methods to be used in the near future to detect and present new words in Estonian.

**Monitoring the Language: Estonian National Corpus**

The Estonian National Corpus is a monitor-type corpus, as it continually expands to include more and more texts over time. It was started in the 1990s and has expanded since then – using different, newly developed collecting methods – to 1.1 billion tokens in the Estonian National Corpus 2017.

In order to monitor language, especially lexical changes, the Institute of the Estonian Language, in cooperation with Lexical Computing Ltd., crawls the Web every two years. The corpora collected, as well as other types of corpora, are used through the corpus query system Sketch Engine (Kilgarriff et al. 2004). The most recent Estonian National Corpus comprises 1.1 billion tokens. It contains semi-automatically collected written texts (203 million tokens), Web texts and the Estonian Wikipedia (up to 2017). The next corpus will be crawled in 2019.

The corpus is available through the Sketch Engine (Kilgarriff et al. 2004) interface. In order to facilitate the corpus-based analysis of Estonian within Sketch Engine, special modules for the Word Sketch (Kallas 2013), Term Extraction (Kallas et al. 2017) and Good Dictionary Example (Koppel 2017; Kosem et al. 2018) functions have been developed. All lexicographic work on contemporary Estonian is based on corpus analysis.

However, the workflow for detecting, monitoring and registering neologisms is not yet automated. Most neologisms are found in a traditional way, described by Kilgarriff et al. (2015) as "reading and marking":

> "lexicographers read texts which are likely to contain neologisms – newspapers, magazines, recent novels – and mark up candidate new words, or new terms, or new meanings of existing words. It is a high-precision, low-recall approach, since the readers will rarely be wrong in their judgments, but cannot read everything, so there are many neologisms that will be missed".

Klosa and Lüngen (2018, 559) also point out that many, but not all, neologisms can be identified by monitoring the language via editorial media evaluation and interpreting the findings on the basis of lexicographic competence. Only automated methods for corpus linguistics can provide a systematic analysis of large amounts of text, offering neologism candidates to lexicographers.

There is an urgent need to set up an infrastructure for neologism detection to supply lexicographers working with neologisms with candidates for inclusion in dictionaries. In the near future we envisage joining or implementing Néoveille, a Web platform for neologism tracking (Cartier 2017). The platform combines state-of-the-art processes to track linguistic changes with a Web platform for linguists to create and manage their corpora, accept or reject automatically identified neologisms, describe linguistically accepted neologisms and follow

their life-cycle on monitor corpora. Néoveille supports the French, Brazilian Portuguese, Chinese, Russian, Czech, Polish and Greek languages.

Next, we will discuss an experimental study for semi-automatic neologism detection carried out in 2018 at the Institute of the Estonian Language to detect new words for our recent explanatory Dictionary of Estonian (DicEst 2019), published in the language portal Sõnaveeb ('Wordweb').

**Detecting New Words: an Experimental Study**

**Theoretical background.** Cartier (2017) distinguishes two types of existing neology tracking system: the Exclusion Dictionary Architecture and Semantic Neology approaches. Exclusion Dictionary Architecture (e.g. Néoveille (Cartier, 2017)[6]) uses the extraction of novel forms from monitor corpora, using lexicographic resources as a reference exclusion dictionary to induce unknown words. Further filters are then applied to eliminate spelling errors and proper nouns. One of the main difficulties is that this method cannot track semantic neologisms (Cartier 2017, 96), i.e. new meanings. As for Semantic Neology, Cartier (2017: 96-97) claims that none of these methods have been exploited in an operational system.

In our experiment, we followed the procedure common to the Exclusion Dictionary Architecture model. The experiment consisted of several stages: (1) the extraction of novel word forms from the Institute's text collection (collected from 2016 to 2018), (2) using available lexicographic resources as a reference exclusion word list to induce unknown words, (3) the filtering of the list (for typographical errors, proper nouns and unassimilated loan words), (4) the compilation of a neologism candidate list, and (5) a lexicographic evaluation of the results.

[6] www.neoveille.org (accessed March 30, 2019)

The experiment was conducted in April 2018. The goal of the experiment was to create a possible neologism detection prototype.

**Compilation of the candidate list.** The initial list of novel word forms for the experiment was generated on the basis of the Institute's text collection (collected from 2016 to 2018). It contained texts from online news outlets (e.g. Äripäev, Õhtuleht, ERR news, raamatupidaja.ee and Arvutimaailm). There was also data from TV subtitles and transcribed books from heliraamat.eki.ee. The data from automatic transcriptions can be noisy and contains a lot of transcribing errors that need to be eliminated. The initial list (712,197 word forms) consisted of word forms that had failed in the automatic morphological analysis.

We used Python 3[7] language and its library, EstNLTK 1.4.1[8], for lemmatization and morphological tagging of the input text. For most of the filtering and sorting, we used the programming language R[9] and its library Tidyverse[10]. Some sorting was also done in Excel, since it enabled us to browse through the data more easily. To filter out unwanted words, we also used regular expressions in our R scripts and in the Notepad++ text editor.

The experiment can be divided into different steps: (1) lemmatization, data selection and the cleaning of selected lemmas, (2) comparing possible lemmas against existing lexicon data, and (3) creating a neologism candidate list. Once these steps are used to create an automatic neologism detection program, we can add a few more functionalities: (4) a way to search and analyze neologisms in a user interface, with usage frequency in a timeline and

[7] https://www.python.org (accessed March 30, 2019)

[8] https://estnltk.github.io/estnltk (accessed March 30, 2019)

[9] https://www.r-project.org/ (accessed March 30, 2019)

[10] https://www.tidyverse.org/ (accessed March 30, 2019)

concordances, and (5) neologism management: accepting, discarding and monitoring words as neologisms.

**Lemmatization, data selection and cleaning.** The dataset used in the experiment originally had 712,197 word forms.

For lemmatization, we created a Python script that used the EstNLTK package suggestions module to go through the given list, deduce possible lemmas and attach part-of-speech tags to them. We grouped these lemmas using the R programming language and got 552,818 lemmas as a result. The results were ordered in Excel. We decided to use only the lemmas that occurred three to 17 times. The higher number was chosen due to the first results being irrelevant and this was where we saw the first possible neologism. We chose the lowest point of three recurring lemmas because we wanted to limit our experiment's dataset, and occurrences below that number didn't seem to yield enough neologism candidates.

The lemmas were cleaned using regular expressions. We removed lines with sequential hyphens and words that had been joined using several hyphens. Since lemmas had part-of-speech tags, we removed all that were detected as names, interjections, pronouns, adpositions, numerals, conjunctions, punctuation forms and verb extensions. We also replaced UTF-16 character sequences with UTF-8 letters, displayed parts of speech with multiple values on separate rows, removed two- or three-letter sequences that were possible appendices left separate because of transcription errors, and removed possible abbreviations and wrote them out in a new file for later study.

After the cleaning step, 5,290 lemmas were left.

**Comparison of lemmas against existing lexicon data.** We wrote an R script to compare the 5,290 lemmas against existing lexicon data. The lexicographic resources used as reference exclusion dictionaries to induce unknown words were: the Explanatory Dictionary of the Estonian Language (EKSS 2009), the Dictionary of Estonian (DicEst

2019), the Dictionary of Foreign Words (VL 2015), the Dictionary of Standard Estonian (ÕS 2013) and the in-house database of new words of the Institute of the Estonian Language.

After the comparison against reference lexicons, 3,722 lemmas were left.

We were also interested in detecting new direct English loanwords. We created a separate R script which compared this list of 3,722 lemmas against the English-Estonian Machine Translation Dictionary (EN-EE). The resulting list was separately compared against Estonian reference lexicon data. After these steps, we arrived at a list of 233 direct English loanword candidates (e.g. *weekend, lite, backup* and *wallet*), to be considered in the future to determine whether some of them should be included in Estonian dictionaries as well. Arleta Adamska-Sałaciak (2016, 758) has pointed out that 'the lines previously drawn regarding the lexicographic treatment of borrowing seem to be shifting': even bilingual dictionaries tend to contain more and more English-based words. It is interesting to note that English is not the only language that words are borrowed from; there were also words from other foreign languages (e.g. *fouetté, laissez-faire, societa* and *bueno*).

**The neologism candidate list.** The list of 3,722 lemmas was cleaned further as it did not seem fully comprehensive. We did a cursory sweep, removing a few remaining proper names; we then looked over the lemmas that had multiple part-of-speech values, leaving only the ones that were correctly tagged, removed a few lemmas with spelling mistakes etc.

After the secondary cleaning step, 2,294 lemmas were left.

**Lexicographic evaluation of the new word candidate list.** The next step was for lexicographers to select words of interest from the 2,294 new word candidate list. This was done manually. An analysis revealed that the list still contained tokenizing errors (e.g. *ganisatsioon* 'ganization'), common spelling mistakes (e.g *aitähh* 'thank you'),

and lemmatization errors (e.g. nouns were left in the genitive and partitive). There were also direct loans from other languages (e.g. *fer-de-lance, fouetté, bordereau, soentjie, societa, bueno* and *laissez-faire*) and from Estonian dialects (e.g. *tüdrik* 'girl', and *mõlemi* 'both'), which hadn't been excluded since the list was compared only to the list of English words and the Estonian written language. Such candidates were discarded as non-words. We estimate that roughly 10% or fewer of the words from the final candidate list were actual neologisms (e.g. *süler* 'laptop' and *akrojooga* 'acrobatic yoga').

Those new word candidates were checked against the Estonian National Corpus 2017 and a decision was taken as to whether they would be included in the DicEst 2019 to be published (e.g. *diakooniline* 'diaconic'), or included in the in-house database of new words (e.g. *baklavaa* 'baklava', *blog* 'blog', *veelkord* 'once more' vs. the standardized lemma forms *baklava, blogi* and the multi-word phrase *veel kord*) for further examination.

The analysis also showed that there were a lot of derivatives (e.g. *digiteerimine* 'digitalizing') and semantically transparent compound words. Traditionally, regular derivatives and semantically transparent compounds have been left out of the headword list of dictionaries due to high regularity, semantic transparency or the restricted capacity of the printed book. This will most likely change in the era of the automatic compilation of dictionaries, as derivatives and compounds act as independent frequent words in a language. Although native speakers might be able to analyze them on the basis of their knowledge of language, L2 learners and speakers would probably need more explicit explanations and usage examples.

One feature that was very typical of the candidate list was the large number of words derived from proper nouns. Around 180 such words were identified (e.g. *lutsiferianism* 'Luciferianism' and *tarsanlik* 'Tarzan-like').

## Registering and Presenting New Estonian Words

**In-house database of new words.** Since 2005 the lexicographers working on modern Estonian dictionaries, both on the descriptive Dictionary of Estonian (DicEst 2019) and on the prescriptive Standard Dictionary of Estonian (ÕS 2013, ÕS 2018), have been collecting candidates for new words in the joint in-house database. Since 2012 we have used the Sketch Engine (Kilgarriff et al. 2014) Word Lists tool. The database contains over 13,000 new words, the majority of which (around 8,000) have been included in different monolingual or bilingual dictionaries. We estimate that about 5,000 words are on the waiting list or have been rejected for inclusion for different reasons, as outlined before.

The average number of new words registered annually has been about 1,500. We have not used the special mark-up for distinguishing neologisms from all other new words, e.g. from older words that have just not been included in the dictionary. In the near future, after moving our different databases to the single database of Ekilex, working on new words and senses will be reorganized.

We will also try to implement crowdsourcing techniques for neologism registration by offering our users the opportunity to propose new words or senses. They can do so by using the feedback forms on our dictionary portals Sõnaveeb ('Wordweb') and e-keelenõu ('e-Language advice').

**Fewer dictionaries, more data: the new database Ekilex and language portal Sõnaveeb ('Wordweb').** Within the framework of the new dictionary-writing system Ekilex (Tavast et al. 2018), we are slowly moving towards something like one huge lexical database of different types of data obtained from many of our dictionaries or further research. It seems more and more appropriate to talk about specific data instead of different dictionaries. Our slogan might be: fewer dictionaries, more data. We should consult on and discuss new words (instead of

having a dictionary of new words) or morphological data (instead of putting this data in some dictionary) or definitions (instead of having the DicEst 2019) or language advice (instead of having ÕS 2018). The new language portal Sõnaveeb (‘Wordweb’) was released in February 2019. As of 27 March, 2019, there are 22,000 users, around 2,000 daily, and the number of sessions per user is 2.90, according to Google Analytics.

The backbone of the database is (and will continue to be) formed from all dictionaries aggregated into the database, with the Dictionary of Estonian (DicEst 2019) constituting the largest part of the database. All new words published in the Sõnaveeb (‘Wordweb’) so far originate from the DicEst (2019), the most recently published lexicographic collection of Estonian.

The Sõnaveeb (‘Wordweb’) contains about 17,000 new words and 1,300 new senses which were either added by lexicographers during the editorial evaluation of print and online media or taken from an in-house database of new words. Neologisms found as a result of our experimental study were also included. All these words and senses were marked as ‘new’ in the original database of DicEst. In this very large number of new words and multi-word expressions, there are also many words which are not at all new in the sense of neologisms. By 'neologisms' we mean words and multi-word expressions that have come into use any time during the last two decades, that denote new phenomena in society and that are perceived by users as new (Langemets et al. 2018).

We estimate that there are about 5,000 neologisms in the Sõnaveeb (‘Wordweb’), e.g. *eelhääletus* ‘pre-election’, *euroskeptitsism* ‘Euroscepticism’, *e-valimised* ‘e-elections’, *kelguhoki* ‘sledge hockey’, and *kendo* ‘kendo’.

Around 1,300 proper nouns were added following the tradition of Indo-European (printed) dictionaries of enriching dictionaries with encyclopedic data: geographic names (e.g. *Beneluxi maad* ‘Benelux countries’ and *Checkpoint Charlie*), well-known fiction or film

characters (e.g. *Ämblikmees* ‘Spiderman’) and names of organizations (e.g. *Taliban*). Proper nouns are described in exactly the same way as any other words. Some new words are derivatives of proper nouns (e.g. *kafkalik* ‘Kafkaesque’) while proper nouns (e.g. *Kafka*) themselves might not be explained as regular entries.

Thousands of ‘new’ multi-word expressions are headwords that have not previously been treated as independent, e.g. *alkohoolne jook* ‘alcohol’, *elu ja surma küsimus* ‘a question of life and death’ and *häälest ära* ‘with no voice’. Also many word forms have lexicalized and registered independently, e.g. *kaelani* (*kaela-ni* ‘neck-TERM’) ‘fully, entirely’. There are a large number of informal words included in the dictionary as well.

New meanings have usually come into being based on similarities with other words, or meanings have widened, influenced by loan words from other languages (e.g. *opereerima* ‘operate in the way equipment does’). In any case, polysemy can be rather universal and is also open to regular semantic transfer, e.g. patterns of systematic polysemy.

**Descriptive vs. prescriptive approach.** Within the single database Ekilex (Tavast et al. 2018), we have to deal with both descriptive and prescriptive data. On the one hand, new words might be seen as exciting new instances in language but, on the other, there is a long tradition of prescriptively pointing out good and bad style in language. There are many controversial cases where data from a descriptive dictionary (e.g. DicEst 2019) is opposed to data from a prescriptive dictionary (e.g. ÕS 2018). Prescriptive data concerns mostly (1) orthography and pronunciation, marking the degree of quantity, stress and palatalization, and (2) inflection (Raadik, Tuulik 2018, 155). The main task is to specify what belongs to standard Estonian and what does not (Raadik, Tuulik 2018, 155). Problems with the descriptive approach arise when dealing with meanings: we think that it is not possible to be prescriptive about meanings, and that it is only possible

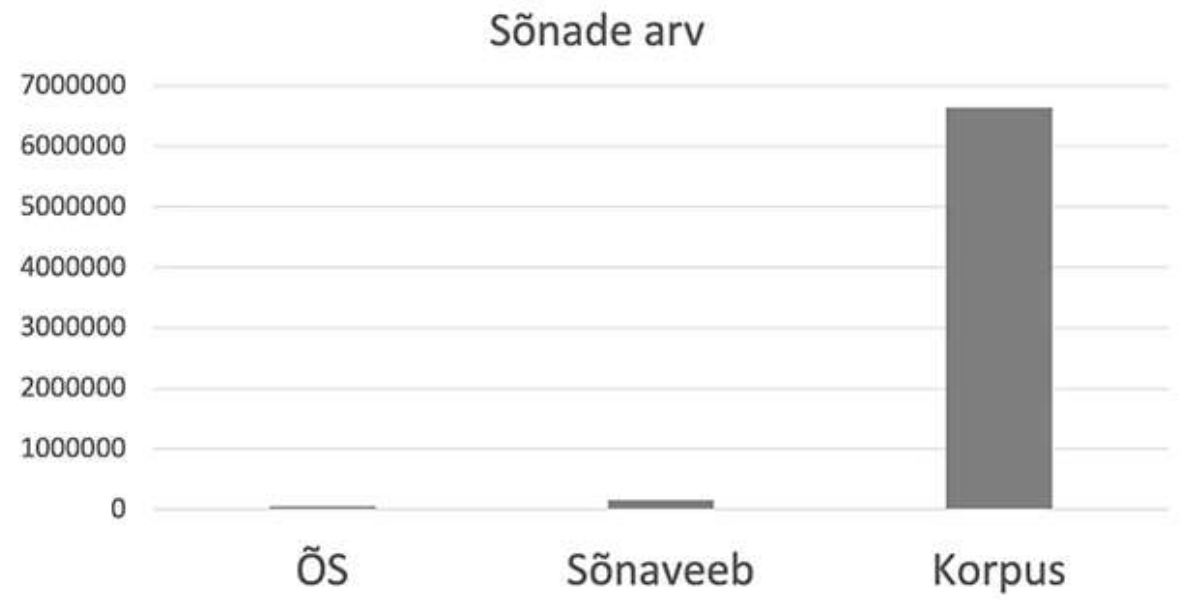
to suggest better writing style, i.e. better wordings for some (over-used) words or phrases, in order to be unambiguous when expressing oneself.

According to government regulations from 2006, the literary norm should be based on the most recent language-planning dictionary issued by the Institute of the Estonian Language, i.e. the printed (!) ÕS 2018. We face the urgent task of arranging data concerning language norms and language advice in a way that is understandable and useful for all kinds of users.

**Conclusions and Future Research**

Tavast (2019) has visualized the real mass of lemmas in the Estonian National Corpus 2017 vs. their actual presentation in dictionaries (Figure 1).

**Figure 1.**




NUMBER OF LEMMAS IN THE CORPUS VS. NUMBER OF LEMMAS IN THE DICTIONARIES. ÕS = Dictionary of Standard Estonian ÕS 2018, Sõnaveeb = language portal Sõnaveeb ‘Wordweb’, and Korpus = Estonian National Corpus 2017

ÕS 2018 contains 54,023 lemmas. The Ekilex database presented by the front-end portal Sõnaveeb (‘Wordweb’) contains 152,978 lemmas. The frequency list of the biggest corpus of Estonian contains 6,637,121 lemmas. Of course, there are numerous non-lexical instances (mistakes, foreign words, proper names, addresses, numbers etc.), as we demonstrated in our experiment, yet it is the only way to search for new words in a growing mass of corpus data. The candidate list provides insights (based on orthography) into signs of deviance in language norms/language change (e.g. *baklavaa* ‘baklava’ et al. mentioned earlier).

As Estonian is a predominantly agglutinative, highly inflected language with a productive and flexible morphological derivation system, none of the dictionaries or central portals could contain ‘all words’ of the language. But it is important to automate the tracking of particular neologisms in order to provide users with up-to-date information on new word meanings and usage. One important question is how to advise users on spelling norms/variants and the word formation of neologisms. In the Institute’s morphological database, we plan to develop a function for automatic morphological paradigm generation in order to be able to generate morphological paradigms for neologisms.

Our experimental study revealed that in order to make neologism discovery more effective we need more advanced tools for automatic language processing. There were a large number of mistakes in tokenization, lemmatization, POS tagging, Name Entity Recognizer (NER) etc. Reference databases need to have not only lists of lemmas from Estonian dictionaries but also lists of lemmas for other languages (mostly English, Finnish, German, Spanish and French) and lists of lemmas generated on the basis of Estonian dialect dictionaries, since there is quite a lot of text on the Web written in Estonian dialects. Also it would be good to create new types of resources, e. g. a database of common spelling mistakes.

Our experimental study was focused mostly on single word detection. Another challenge is the detection of multi-word expressions and new meanings. There is a clear need for the implementing of Semantic Neology methods (Cartier 2017, 96-97). So far, no research has been conducted in this field for Estonian.

To track neologisms we envisage joining or implementing Néoveille, a Web platform for neologism tracking (Cartier 2017). The platform combines state-of-the-art processes to track linguistic changes with a Web platform for linguists to create and manage their corpora, accept or reject automatically identified neologisms, describe linguistically the accepted neologisms and follow their life-cycle on monitor corpora.

When presenting lexical information in the language portal Sõnaveeb ('Wordweb'), we plan to visualize usage and frequency information on the basis of time-stamped corpora. At the moment, there are no time-stamped corpora for Estonian, but there is a huge need for this resource.

Within the unified single database Ekilex, we will present both descriptive and prescriptive data. We hope to make the system as flexible as possible to satisfy the needs of all kinds of users.

## References

Aavik, Johannes. 1919. *Uute sõnade sõnastik. [Dictionary of new words.]* Tartu: Istandik, 1919. (2. edition 1921"Uute ja vähem tuntud sõnade sõnastik").

Adamska-Sałaciak, Arleta. 2016. On bullying, mobbing (and harassment) in English and Polish: Foreign-language-based Lexical Innovation in a Bilingual Dictionary. In *Proceedings of the 17th EURALEX International Congress, 2016.* Eds. Tinatin Margalitadze, George Meladze. Tbilisi, Georgia. Ivane Javakhishvili Tbilisi University Press, 758-766.

Cartier, Emmanuel. 2017. Néoveille, a Web Platform for Neologism Tracking. In *Proceedings of the EACL 2017 Software Demonstrations*, Valencia, Spain, April 3-7 2017. 95–98.

DicEst 2019 = *Eesti keele sõnaraamat 2019. [The Dictionary of Estonian 2019.]* Eds. Margit Langemets, Mai Tiits, Udo Uibo, Tiia Valdre, Piret Voll. Compiled by Katrin Kuusik, Külli Kuusk, Margit Langemets, Mai Tiits, Udo Uibo, Tiia Valdre, Piret Voll. Eesti Keele Instituut. Sõnaveeb 2019. https://sonaveeb.ee (accessed March 30, 2019).

EKSS 2009 = *Eesti keele seletav sõnaraamat I–VI. ("Eesti kirjakeele seletussõnaraamatu" 2., täiendatud ja parandatud trükk.) [The Explanatory Dictionary of Estonian.]* Eds. Margit Langemets, Mai Tiits, Tiia Valdre, Leidi Veskis, Ülle Viks, Piret Voll. Eesti Keele Instituut. Tallinn: Eesti Keele Sihtasutus, 2009. http://www.eki.ee/dict/ekss/ (accessed March 30, 2019).

EN-EE = *Inglise-eesti masintõlkesõnastik. [English-Estonian Machine Translation Dictionary.]* http://www.eki.ee/dict/ies/index.cgi (accessed March 30, 2019).

Erelt, Tiiu, Meriste, Huno, and Rein Kull. 1985. *Uudis ja unarsõnu. [New and forgotten words.]* Tallinn: Valgus, 1985.

Grenzstein, Ado. 1884. *Eesti sõnaraamat. [Estonian dictionary.]* Tartu: Oma trükk ja kirjastus, 1884.

Kilgarriff, Adam, Rychly, Pavel., Smrž, Pavel and David Tugwell. 2004. The Sketch Engine. *In Proceedings of the XI Euralex International Congress.* Eds. Williams G. and Vessier, S. Lorient: Université de Bretagne Sud, 105–116.

Kilgarriff, Adam, Bušta, Jan and Pavel Rychlý. 2015. DIACRAN: a framework for diachronic analysis. *Corpus Linguistics (CL2015)*, United Kingdom, July 2015. https://www.sketchengine.eu/user-guide/user-manual/trends/#toggle-id-6 (accessed March 30, 2019).

Klosa, Annette, and Harald, Lüngen. 2018. New German Words: Detection and Description. In *Proceedings of the 18th EURALEX International Congress: Lexicography in Global Contexts.* Eds. Čibej, J. Gorjanc, V., Kosem, I. and Krek, S. Ljubljana, Slovenia 17-21 July 2018, 559−569. http://euralex2018.cjvt.si/publication/2 (accessed March 30, 2019).

Raadik, Maire and Maria, Tuulik. 2018. Borrowed material as approached by Estonian language planning practitioners: The experience of the Dictionary of Standard Estonian. In *International Journal of Lexicography*, Vol. 31, No. 2, June 2018, 151–166.

Sõnaveeb = *Sõnaveeb 2019. [The Wordweb 2019.]* Developed and edited by Indrek Hein, Jelena Kallas, Kristina Koppel, Margit Langemets, Kaur Männiko, Tõnis Nurk, Ülle Viks. Developer OÜ TripleDev: Martin Laubre, Raigo Ukkivi, Arvi Tavast, Sander Lastovets, Sander Rautam. Eesti Keele Instituut. Sõnaveeb 2019. http://www.sonaveeb.ee (accessed March 30, 2019).

Tavast, Arvi, Langemets, Margit, Kallas, Jelena and Krsitina Koppel. 2018. Unified data modelling for presenting lexical data: The Case of EKILEX. In *Proceedings of the XVIII EURALEX International Congress. EURALEX: Lexicography in Global Contexts*. Eds. Čibej, J., Gorjanc, V., Kosem, I. and Krek, S. Ljubljana, 17–21 July 2018. Ljubljana: Ljubljana University Press, Faculty of Arts, 749−761.

Tavast, Arvi. 2019. *1000 sagedamat sõna, mida ÕSis pole. [The 1000 most frequent words missing from Dictionary of Standard Estonian ÕS 2018]*. Blog posted on February 8, 2019 by Arvi Tavast. http://tavast.ee/1000-sagedamat-sona-mida-qsis-pole/ (accessed March 30, 2019).

ÕS 2013 = *Õigekeelussõnaraamat ÕS 2013. [The Dictionary of Standard Estonian ÕS 2013]* Ed. Maire Raadik. Tallinn: Eesti Keele Sihtasutus, 2013. http://www.eki.ee/dict/qs2013/ (accessed March 30, 2019).

ÕS 2018 = *Eesti õigekeelsussõnaraamat ÕS 2018. [The Dictionary of Standard Estonian ÕS 2018]* Ed. Maire Raadik. Compiled by Tiiu Erelt, Tiina Leemets, Sirje Mäearu, Maire Raadik. Eesti Keele Instituut. Tallinn: Emakeele Sihtasutus, 2018). http://www.eki.ee/dict/qs2018/ (accessed March 30, 2019).

VL 2015 = *Võõrsõnade leksikoni veebiversioon. [The Web Dictionary of Foreign Words.]* Ed. Tiina Paet. Eesti Keele Instituut. http://www.eki.ee/dict/vsl/ (accessed March 30, 2019).