Jelena Kallas/Margit Langemets/Tõnu Tender

Opening up Estonian dictionaries to European communities and language technology

Abstract

This paper describes recent developments at the Institute of the Estonian Language concerning e-lexicography in order to open up lexical resources created by the Institute to European communities and language technology.

We describe the main public dictionary services created by the Institute. Within the framework of the new dictionary writing system Ekilex, the Institute is moving away from presenting separate interfaces for different dictionaries towards a unified data model in order to provide data in an aggregated form via the language portal Sõnaveeb. For industrial partners, data can be accessed via Ekilex API. We also give a short overview of the Ekilex data model and methodology applied to the unification of dictionaries. Finally, we describe the accessibility policy and availability of language resources created by the Institute and propose development perspectives.

1. Public dictionary services offered by the Institute

The Institute of the Estonian Language[1] (henceforth the Institute) is a national research and development institution funded by the Ministry of Education and Science. It performs a number of public functions: (a) compiling and upgrading dictionaries and databases essential to the country and national culture, including Estonian dialects and Finno-Ugric languages, (b) language care and language planning, (c) coordinating nationwide terminological work, (d) providing services for language learners and people with special needs and (e) developing speech synthesis for Estonian. Its long-term operation as an independent institution has brought the Institute a nationwide reputation as the centre of Estonian studies (see the Corporate Image Survey by TNS Emor 2015). The name of the Institute of the Estonian Language on title pages and on the web has become a symbol of quality.

Practical language usage in Estonia is regulated by the Law on Language and legislation based on that. According to a government regulation in 2006, the literary norm should be based on the most recent Dictionary of Standard Estonian issued by the Institute of the Estonian Language. The new edition of the dictionary thus became the "updated" official norm of the standard language (Langemets 2013).

[1] https://eki.ee/ (last accessed April 4, 2020).

Since the autumn of 2015, the Institute of the Estonian Language has been the only compiler of major academic dictionaries in Estonia. The Institute owns and hosts more than 100 dictionaries and terminological databases, which altogether contain about 1.5 million words or terms. Users make about 7.8 million queries per year on the Institute's website. The language hotline "e-keelenõu"[2] receives about 100,000 queries per month.

Out of the numerous public functions performed by the Institute, we will next focus on lexicographic tools and services in progress:

– building the new Dictionary Writing System Ekilex[3] (since 2017),
– providing content for users via the new language portal Sõnaveeb[4] [Wordweb] (since 2019),
– opening up terminological databases (since 2020),
– accessibility policy.

2. Building the new Dictionary Writing System Ekilex

The last decade has been a transitional period in Estonia, when dictionaries were published on paper as well as electronically. So far the majority of online versions have been almost exact copies of paper dictionaries. 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 lexical data in an aggregated form for the user via the new language portal Sõnaveeb. Ekilex is maintained and developed by the Institute in collaboration with the software company TripleDev.

The conceptual design of a comprehensive lexical resource must be based on a theoretical understanding of the organisation of linguistic knowledge. While the traditional formats of language description are mostly based on a modular view of language, recent decades have seen the rapid theoretical development of a non-modular, usage-based conception of language.

The data model of Ekilex and fundamental issues connected with the creation of the unified database were described in Tavast et al. (2018) and Koppel et al. (2019). Here we outline the most important aspects.

The data model has many-to-many relations between words and meanings and is suitable for both word-based and concept-based representations of data. When importing new data from different datasets, we try to keep words and meanings as dataset-agnostic units, allowing a gradual transition from the initial condition of several independent datasets to the end goal of a single Ekilex resource containing all lexical information known about Estonian. We only import pieces of information

[2] http://keeleabi.eki.ee/ (last accessed April 4, 2020).
[3] https://ekilex.eki.ee/ (last accessed April 4, 2020).
[4] https://sonaveeb.ee/ (last accessed April 4, 2020).

that clearly add value and do not duplicate the data already presented in the database. This means that when moving towards a single database, added datasets are turned into information layers and applied to the central "backbone" of headwords already present in the database, removing the need to specify variations of the same information again in separate dictionaries.

The initial import of separate lexical datasets resulted in a massive duplication of both words and meanings, and it is taking a great effort to harmonise the data by 1) unifying homonyms (the work is partly done automatically) and 2) unifying meanings (the work is mostly done manually). The migration of data from separate datasets to a single resource has also eliminated the need to harmonise the presentation of the same pieces of information (e.g. domain and register labelling) in the database.

The current descriptions of Estonian items in Ekilex include sense definitions, semantic types, word classes, inflectional forms, collocations, government patterns, semantic relations, related words, etymology and usage examples, including automatically retrieved corpus examples, pronunciations of basic word forms and usage examples, and corpus frequency.

In the near future, we foresee extending the scope of Ekilex to representing prescriptive data, grammar description (as schematic constructions), bilingual data (as there are more bilingual databases available at the Institute, e.g. Latvian, Finnish and Chinese) and information on language proficiency levels corresponding to the Common European Framework of Reference for Language (CEFR 2001) and its companion volume with new descriptors (CEFR/CV 2018). Prescriptive data will constitute a major change in the present Dictionary of Standard Estonian ÕS 2018.[5]

3. The new language portal Sõnaveeb [WordWeb]

Sõnaveeb [WordWeb] is the Institute's language portal containing linguistic information from a growing number of dictionaries and databases (Koppel et al. 2019). The portal was released in February 2019. The information displayed in Sõnaveeb comes from Ekilex. As of February 2020, Ekilex contains about 70 lexical databases, with both general and specialised dictionaries. Databases are constantly updated and edited, including changes that are made upon receiving feedback from users. As of February 2020, the portal contained about 170,000 words and phrases in Estonian, about 70,000 words and phrases in Russian and 40,000 words and phrases in English. The versions of Sõnaveeb are updated and archived once a year.

Beginning in 2020 all information from separate databases will be displayed in a unified mode as a single source called EKI ühendsõnastik [EKI Combined Dictionary]. The combined dictionary for Estonian displays information from

[5] http://www.eki.ee/dict/qs/ (last accessed April 4, 2020).

different lexical databases: *The Dictionary of Estonian* (2019), the *Estonian Collocations Dictionary* (2019), the *Basic Estonian Dictionary* (2014) and *The Estonian Morphological Database of the Institute of the Estonian Language* (2019). It also displays information from bilingual lexical databases: the *Estonian-Russian Orthographic Dictionary for Students* (2018; 1st edition 2011) and the *Estonian-Russian Dictionary* (2018; 1st edition 1997-2009).

In addition to carefully selected usage examples in the EKI Combined Dictionary, we display web examples from *The Corpus of Web Examples for Estonian* (2020) via the corpus query system KORP API.

The versions of the Sõnaveeb portal and the EKI Combined Dictionary are updated every month and archived once a year.

All lexicographic work on contemporary Estonian is based on corpus analysis and we provide the links to corpora when presenting dictionary content via the Institute's new language portal Sõnaveeb. The biggest corpus of Estonian at the moment is the (2019) Estonian National Corpus (1.8 billion tokens). It is available through the Sketch Engine[6] interface (Kilgarriff et al. 2004).

4. Opening up terminological databases

Since March 2020, Sõnaveeb displays information from over 60 terminological databases. The user searching for a term is provided with data on both general language and specialised language. The presentation mode for terminological data will be tuned to enable both a word-based view (as in Sõnaveeb) and a concept-based view (in the near future). There are still many other (smaller) termbases that will be added to Ekilex in the future.

The biggest termbase in Ekilex is the multilingual termbase Esterm containing over 50,000 concepts and 150,000 terms in five languages. The second biggest database is the defence and military database Militerm, containing approx. 4,000 concepts.

The Ministry of Education and Science launched the national programme for developing terminological work in 2008. One of its aims was to build a unified system for providing terminological knowledge to the public. This has taken a good deal of time. Finally we have succeeded in opening up terminological resources to the public.

5. Accessibility policy and availability of language resources for Estonian

All electronic resources and applications developed by the Institute are available to the public for free.

[6] www.sketchengine.eu/ (last accessed April 4, 2020).

On the Institute's website, online dictionaries and databases can be found, as well as corpora, language and speech technology applications, termbases, etc. Some data from the dictionaries are freely available for download, mostly under the terms and conditions of the Creative Commons BY 4.0 licence.

The Exilex data can be used by external (industrial) partners via API. As mentioned before, in the Ekilex data model, words (i.e. headwords) and meanings (i.e. definitions and domain indicators) are dataset-agnostic. After having processed, systematised, unified, supplemented, edited, etc. the information across datasets, the Ekilex resource has attained the status of a single database called the EKI Combined Dictionary, licensed under the Creative Commons BY 4.0 International license. Data containing personal data and third-party data with reusability restrictions are licensed under CLARIN Academic EULA v1.0, including appropriate additional terms (identification, access, general use and distribution conditions). The metadata on created resources are findable in the META-SHARE repository.[7] Where applicable, the Institute follows the recommendation on legal and intellectual property rights issues for lexicography (Boelhouwer et al. 2020) outlined in the Horizon 2020 project European Lexicographic Infrastructure.

6. Conclusion

The future work of the Institute is strongly connected with the Institute's new Dictionary Writing System Ekilex (Tavast et al. 2018) and the Ekilex-based language portal Sõnaveeb. The long-term vision is to have a single data source (Ekilex) that provides (also via the API) consistent and comprehensive information about Estonian words, combining the research done in all departments and working groups of the Institute.

Undoubtedly, there will be more exciting challenges in the near future as we continue shifting from compiling stand-alone dictionaries with incompatible data structures to integrating lexicographic data into a unified and standardised database, and making the data findable, accessible, interoperable and reusable. These issues are very much in line with the objectives and outcomes of the Horizon 2020 project European Lexicographic Infrastructure (ELEXIS),[8] developing strategies for structuring and linking lexicographic resources. The Institute will follow standards developed within the ELEXIS project in order to keep Estonian dictionaries open to European communities and language technology fields, such as Natural Language Processing, Linked Open Data and the Semantic Web.

[7] https://metashare.ut.ee/ (last accessed April 4, 2020).

[8] https://elex.is/ (last accessed April 4, 2020).

7. Acknowledgements

This work has been partially supported by funding from the European Regional Development Fund.

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 731015.

References

Boelhouwer, B./Kosem, I./Nimb, S./Jakubíček, M./Tiberius, C./Krek, S./Rosenmeier M. *(2020): Recommendations on legal and IPR issues for lexicography. Deliverable of ELEXIS project*. https://elex.is/wp-content/uploads/2020/02/ELEXIS_D6_2_Reccommendations_on_Legal_and_IPR_Issues_for_Lexicography.pdf (last accessed March 8, 2020).

CEFR 2001: *The Common European Framework of Reference for Languages: Learning, teaching, assessment*. Strasbourg: Council of Europe.

CEFR/CV 2018: *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume with new descriptors*. Strasbourg: Council of Europe

Kilgarriff, A./Rychlý, P./Smr, P./Tugwell, D. (2004): The Sketch Engine. In: Williams, G./Vessier, S. (eds.): Proceedings of the 11th EURALEX International Congress. Lorient: Université de Bretagne Sud, 105-115.

Kallas, J./Langemets, M./Koppel, K./Tuulik, M. (2019): State-of-the-art on monolingual lexicography for Estonia. In: *Slovenščina* 2.0, 7(1), 25-38.

Koppel, K./Tavast, A.Langemets, M./Kallas, J. (2019): Aggregating dictionaries into the language portal Sõnaveeb: Issues with and without solutions. In: Kosem, I./Zingano Kuhn, T./Correia, M./Ferreria, J.P./Jansen, M./Pereira, I./Kallas, J./Jakubíček, M./Krek, S./Tiberius, C. (eds.): *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 Conference. 1-3 October 2019, Sintra, Portugal*. Brno: Lexical Computing CZ, 434-452.

Langemets, M. (2013): 'To think outside the paper'. The case of Estonia. In: Stickel, G./Váradi, T. (eds.): Lexical challenges in a multilingual Europe. Contributions to the Annual Conference 2012 of EFNIL in Budapest. (= Duisburger Arbeiten zur Sprach- und Kulturwissenschaft 99). Frankfurt a.M./Berlin: Peter Lang, 145-161.

Tavast, A./Langemets, M./Kallas, J./Koppel, K. (2018): Unified data modelling for presenting lexical data: The case of EKILEX. In: Čibej, J./Gorjanc, V./Kosem, I./Krek, S. (eds.): *Lexicography in global contexts. Proceedings of the XVIII EURALEX International Congress, Ljubljana, 17-21 July 2018*. Ljubljana University Press, Faculty of Arts, 749-761.

TNS Emor (2015): *Eesti Keele Instituudi tuntuse ja maine uuring 2015* [The Corporate Image Survey]. http://portaal.eki.ee/dokumendid/category/10-infoks.html.