

From online dictionaries to language infrastructure: developments in Estonia



Kristina Koppel, Senior Computational Lexicographer
Institute of the Estonian Language | EKI
kristina.koppel@eki.ee

” Overview

- Current state of modern lexicography
- DWS's used in the Institute of the Estonian Language (IEL)
- Main components of Estonian language infrastructure:
 - **Ekilex** – a new dictionary writing system
 - **Sõnaveeb** – a new language portal
 - **EKI Combined Dictionary** – largest (general) dictionary of modern Estonian
 - **EKI Reference Book** – collection of texts serving for information and advice
- Tools for learners and teachers of L2 Estonian:
 - **Keeleõppija Sõnaveeb**
 - **Teacher Tools**
 - **Picture Dictionary**

Modern lexicography

The state of the art in lexicography has evolved from

- paper to electronic
- introspective to empirical
- manual writing to corpus based generation
- normative to descriptive
- binary to quantitative
- human only to machine-readable

Focus has shifted from compiling and publishing stand-alone dictionaries to keeping the information about words in **one database** so that the data would be findable, accessible, interoperable and reusable

Lexicographers need to pay attention to quality of lexicographic data and data modeling of lexicographic databases



Dictionary and termbases in IEL

IEL has been publishing dictionaries and termbases for decades, e.g. descriptive and normative dictionaries, monolingual and bilingual dictionaries, dictionaries for L2 learners, collocation and synonym dictionary, the dictionary of Estonian dialects, the dictionary of word families, etc.

A high degree of autonomy: separate databases, separate webpages

DWS's used in IEL

1. EELex (2003–2015)

- currently holds over 70 dictionary databases of different types
- started as a XML-database, later transferred to a mixed model storing chunks of XML in a relational database
- semasiological data model

2. Termeki (2007–2015) by Werkdata Ltd. (termbases.eu)

- free to Estonian terminologists
- mainly used outside of the IEL
- onomasiological data with relational database

3. Multiterm

- used for two major termbases at the IEL
- onomasiological data model with XML-database

 EELex

Very flexible – a new data model for each new dictionary

- accommodated the heterogenous wishes of lexicographers
- each author obtained a data model of their choice

Results were not in line with current thinking in lexicography

- datasets were disconnected
- information was duplicated and inconsistent across datasets
- the same information was located differently in the model depending on the dataset

Online publishing: each dataset had a separate public interface

→ Changes in working methods and tools were inevitable



Ekilex

”**Ekilex** (2017–...)

A new DWS for both semasiological dictionaries as well as onomasiological termbases

A single data resource that provides consistent information about Estonian words

Copes with the multitude of existing datasets, allows importing

Data model is based on a m:n relation between words and meanings

Tavast, Arvi; Langemets, Margit; Kallas, Jelena; Koppel, Kristina (2018). Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. Proceedings of the XVIII EURALEX International Congress: EURALEX: Lexicography in Global Contexts, Ljubljana, 17-21 July 2018. Ed. Jaka Čibej, Vojko Gorjanc, Iztok Kosem, Simon Krek. Ljubljana University Press, Faculty of Arts, 749–761.

- Kokku leiti: 2
- 1 auto (et)
eki, les, mil
Avalik süno valmis
 - 2 auto- (et)
õtb

auto Ühenda Märki tehtuks 28.07.2023
eesti + Vormikood + Häälendus auto + Sugu + Keelendi tüüp + Aspekt

1 sün + automobiil van [1] | neljarattaline [0.7] | masin kõnek [0.9] | tõld kõnek [0.7] | sõiduauto [0.7] | trandulett [0.9] | käru kõnek [0.9] | sõiduk, sõiduriist, sõiduvahend [0.7] | ratsu [0.7] | jõuvanker van [0.9]

en car [1]
fr voiture [1]
ru автомобиль [1] | машина [1] | автомашина [1]
uk автомобіль [1] | автомашина [1] | машина [1]

Ühenda Duplikaat Uus keelend Uus tähendus Kustuta Logi Sisemärkus

IDENTIFIKAATORID W159100 L1151531 M15403
ILMIKU SILT + süno valmis , prantsuse tehtud
DETAILSUS Lihtne/Detailne
KEELETASE A1
ILMIKU KAAL 1
SÕNALIIK + s - nimisõna, substantiiv
SEMANTILINE TÜÜP + ese_instru - tööriist, sõiduvahend jt

SELETUS + et
• sõitjate või veose veoks mõeldud vähemalt neljarattaline mootorsõiduk [eki]
• mootoriga sõiduk, millel on rattad all ja millega inimesed saavad liikuda ühest kohast teise [eki]

PILT +



- ÖS selgitab
Sisemärkus
Sõnavormid
16 noomen
auto SgN - ainsuse nimetav
auto SgG - ainsuse omastav
auto[t SgP - ainsuse osastav
- SgAdt - ainsuse suunduv e lühike sisseütlev
auto[sse SgIII - ainsuse sisseütlev
auto[s SgIn - ainsuse seesütlev
auto[st SgEI - ainsuse seestütlev
auto[le SgAll - ainsuse alaleütlev
auto[l SgAd - ainsuse alalütlev
auto[lt SgAbI - ainsuse alaltütlev
auto[ks SgTr - ainsuse saav
auto[ni SgTer - ainsuse rajav
auto[na SgEs - ainsuse olev
auto[ta SgAb - ainsuse ilmaütlev
auto[ga SgKom - ainsuse kaasaütlev
auto[d PIN - mitmuse nimetav
auto[de PIG - mitmuse omastav
auto[sid PIP - mitmuse osastav
auto[desse PIII - mitmuse sisseütlev
auto[des PIIn - mitmuse seesütlev
auto[dest PIEI - mitmuse seestütlev
auto[dele PIAII - mitmuse alaleütlev
auto[del PIAAd - mitmuse alalütlev
auto[delt PIAAbI - mitmuse alaltütlev
auto[deks PITr - mitmuse saav
auto[deni PITer - mitmuse rajav
auto[dena PIEs - mitmuse olev
auto[deta PIAAb - mitmuse ilmaütlev
auto[dega PIKom - mitmuse kaasaütlev
- Sõna seos
Ühendid
Veel sarnaseid sõnu

From dictionaries into information layers

Combining legacy dictionaries into a single data source (EKI Combined Dictionary):

1. Dictionary of Estonian 2019
 - descriptive corpus-based comprehensive scholarly dictionary
 - focused on written Estonian
 - describes senses, hence serves as a **database backbone**
2. Estonian Collocations Dictionary 2019
3. Basic Estonian Dictionary 2019
4. Estonian-Russian Dictionary
5. MAB (morphophonological datasets)

All datasets or dictionaries are turned into information layers and applied to the central backbone, which then removes the need to specify variations of the same information again in separate dictionaries



EKI Combined Dictionary

The largest (general) dictionary of modern Estonian (~170 000 Estonian headwords)

Datasets are constantly updated and edited by ~30 lexicographers, including changes that are made upon receiving feedback from users

Contains multiple information layers (definitions, morphology, neologisms, examples, collocations, synonyms, equivalents, etymology, normative recommendations, government, MWEs, language proficiency levels etc.)

”Termbases

Since 2019, IEL coordinates terminology work in Estonia

120+ termbases, including e.g. genetics, geology, military, robotics, tourism, religion

At the moment they are all separate databases



Sõnaveeb

Language portal Sõnaveeb

Consists data from a growing number of dictionaries and termbases (in Ekilex)

Serves human users as an aggregator with items of content to one web page and enables access to data within several dictionaries

Currently holds a total of 285 000 Estonian (212 000 Russian / 94 000 English, ...), obtained from 124 databases

80 000 users per month, 55% mobile / 45% desktop

The screenshot shows the homepage of the Sõnaveeb language portal. At the top center is the logo, a purple circle with a white 'S' shape, followed by the text 'Sõnaveeb' and 'Eesti Keele Instituut'. Below the logo is a search bar with the placeholder text 'Search Sõnaveeb'. To the right of the search bar are icons for keyboard input, voice search, and a magnifying glass. Below the search bar are two links: 'Learner's Sõnaveeb' with an external link icon and 'Feeling lucky'. At the bottom of the page, there are several logos: the Eesti Keele Instituut logo on the left, followed by two logos for the European Union (Euroopa Liit) and the Estonian Future Fund (Eesti tuleviku heaks), and the logo of the Ministry of Education and Research (HARIDUS- JA TEADUSMINISTEERIUM) on the right.

www.sonaveeb.ee

Sõnaveeb

Primarily for native speakers

Displays both general and specialized language

Language All languages ▾ Databases All databases ▾ Feeling lucky Feedback

et haug substantive 14.04.2023

EKI COMBINED DICTIONARY 2023

1 **et** põhjapoolkera mage- ja riimvete tugeva pea ja pika sihvaka kehaga röövkala

Synonyms **havi**, havipurikas INFORMAL, haugipurikas, purikas

en pike
la *Esox lucius*
fr brochet
ru щука
uk щука

Usage examples
Üks haug läks mul landi otsast minema. 🔊

Collocations
suur **haug** | kilone | alamõduline ...

1.1 **et** haugiliha toiduna

ru щука

Usage examples
Küpsetatud haug koorekastmes. 🔊

TERMINOLOGICAL DATABASES

Kokanduse terminibaas
ID: 464065 11.08.2023
mereannid

et haug
substantive

en pike
substantive

Word forms 📄

Inflection Type 22e 🔗

haug 🔊	haugid
haugi 🔊	haugide
haugi 🔊	hauge ~ haugisid

🔗 Show as table

Phrases and phrasal verbs 📄

haugi mälu

More similar words 📄

havinolk, pulk, purikas, vaaphaug, jäähaug
...

Possible translations 📄

ru щука

Web examples 📄

⚠️ These examples have been automatically selected and may contain errors.

Järves leidub **haugi**, särge ja linaskit.

Järves leidub tänini ahvenat, forelli ja **haugi**.

...

Examples from Estonian-Russian dictionary (2019) 📄

harilik haug - обыкновенная щука
must haug - полосатая щука

EKI Reference Book (EKI teatmik)

A collection of texts serving for information and advice (e.g. the approved orthographic rules)

EKI Combined Dictionary and EKI reference book are the main components of Estonian language infrastructure

10 000 users per month

Interlinked with Sõnaveeb

eki.ee/teatmik

ÕS SELGITAB

Info Selgitused on töös, nende sõnastus võib muutuda. Järgmine ÕS ilmub 2025. aastal.

Sõna **diisel** kasutus tähenduses 'diislikütus' on eesti keele ühendkorpuse (2019) järgi tänapäeval üldkeeles levinud. Erialakeeles võib kehtida teisi kokkuleppeid, vt näiteks diislikütus oskussõnastikes. Vt ka 2021. a uurimust ning EKI teatmiku artikleid „Võõrsõnade tähendused“ ja „Tähenduste normimisest“.

(25.01.2023)



Tools for learners and teachers of Estonian L2

Learner's Sõnaveeb

Primarily for learners at the A2–B1 proficiency levels

6,500 basic Estonian words

Shorter definitions, controlled vocabulary, CEFR levels, explicit information about the most frequent morphological forms, pictures, pronunciation

The screenshot shows the dictionary entry for the Estonian word 'haug'. The interface includes a search bar at the top with the word 'haug' entered. The main content area displays the word 'haug' with its part of speech 'substantive' and a date '14.04.2023'. Below this, there is a definition in Estonian: 'suur kala, kes sööb väikseid kalu ja elab ka Eestis'. To the left of the definition are language codes and their corresponding words: 'en pike', 'fr brochet', 'ru щúка', and 'uk щúка'. Underneath the definition are 'Usage examples' with audio icons: 'Kalamees sai suure haugi.' and 'Värske haug on väga maitsev.'. A detailed illustration of a pike fish is shown below the examples. On the right side of the page, there are several sections: 'Word forms' showing inflection types and forms like 'haug', 'haugi', and 'haugisid'; 'More similar words' listing 'pulk'; 'Possible translations' showing 'щúка' in Russian; 'Web examples' with a warning icon and a sentence in Estonian: '"Eesti jõgedes leiduvatest kaladest on enam levinud haug, jõeforell, ahven ja koha."'; and 'Examples from Estonian-Russian dictionary (2019)' with two entries: 'harilik haug - обыкновенная щука' and 'must haug - полосатая щука'.

www.sonaveeb.ee/lite

Teacher Tools

Toolbox for L2 teachers and specialists

Four modules:

1. **Vocabulary** – covers both young (pre A1-B2) and adult (A1-C1) learners
2. **Grammar** – covers only young learners (pre A1-B2)
3. **Language use situations** – offers information about the typical situations where learner should be able to communicate
4. **Text evaluation** – runs on morphological analyser and marks lemmas in texts according to their CEFR-assignment in vocabulary and grammar profile

Keeleoskustase: eelA1 A1 A2 B1 B2 C1

Eesti keel on emakeeleks umbes miljonile inimesele üle maailma ja üha rohkem õpitakse seda ka teise või võõrkeelena. Arvatakse, et eesti keelt on raske õppida, kuna selles on koguni 14 käänet, mis tundub jube hirmutav. Küsimusi tekkitab ka eesti keele sõnajärg. Nagu teada, ei ole eestikeelse lause sõnade järjekord kindlate reeglitega määratud, nagu on seda paljudes indo-euroopa keeltes. Kindlasti on see arvutile seni veel ületamatu raskus ning võib olla keeruline ka eesti keele õppijate jaoks. Näiteks kui vaadata kahte lauset « Lapsed sõid need kommid ära » ja « Need kommid sõid lapsed ära », siis nendes lausetes saab alust ja sihitist määrata ainult lause tähendust teades. Olgem siis lausete tähenduse mõistmisel tähelepanelikud!

Tase	Sõnu	%
eelA1	67	61.47%
A1	11	10.09%
A2	10	9.17%
B1	11	10.09%
B2	4	3.67%
määramata	6	5.50%

Teksti loetavuse hindajad®
Lix-indeks ④: 47 - raske lugeda
Formaalsusindeks ④: 49.1 - keskmise väljendustäpsusega
Nominaalsus ④: 29.4% - tavapärase hulk nimisõnu
Nimi- ja tegusõnade suhe ④: 59 : 41

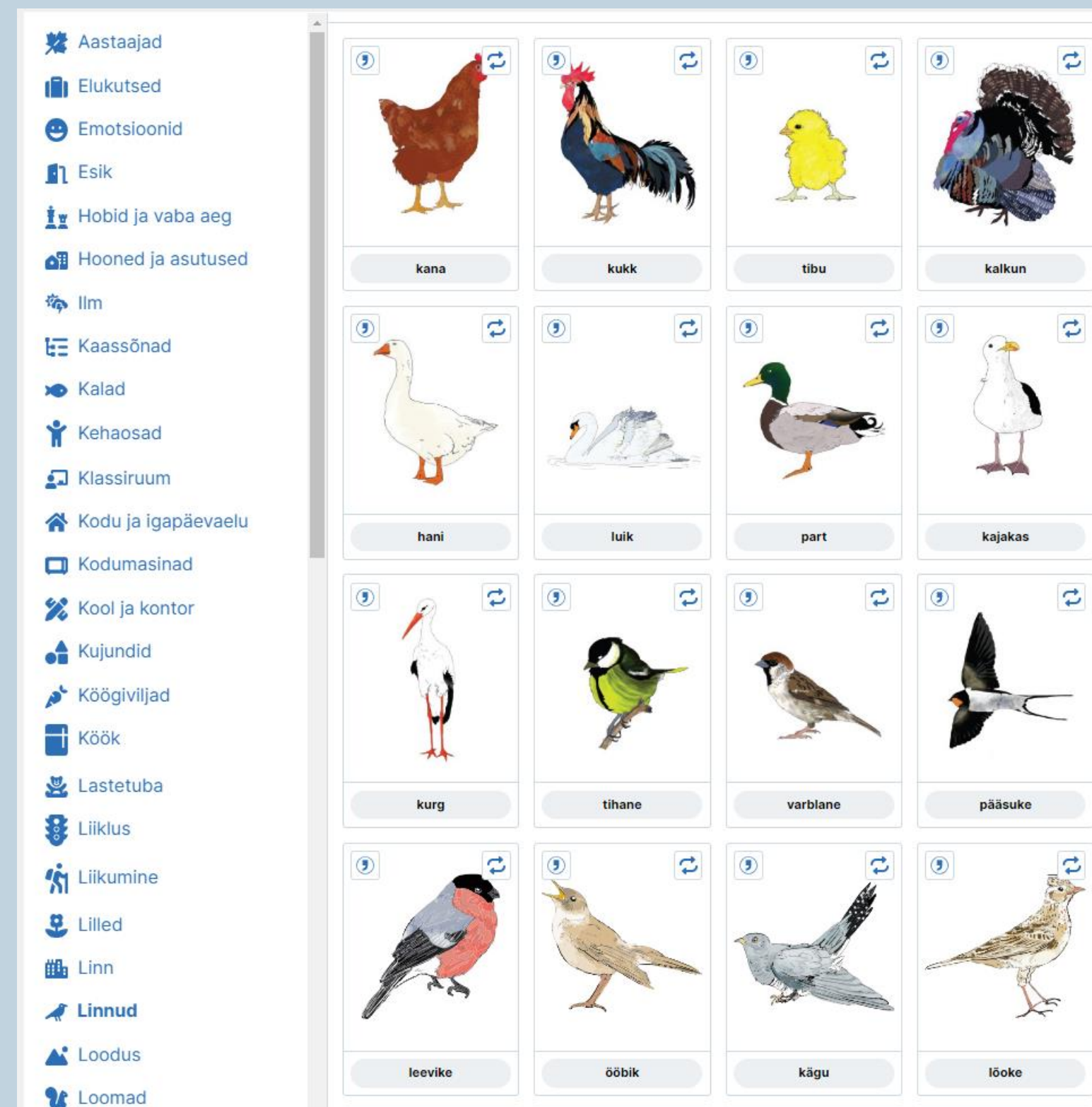
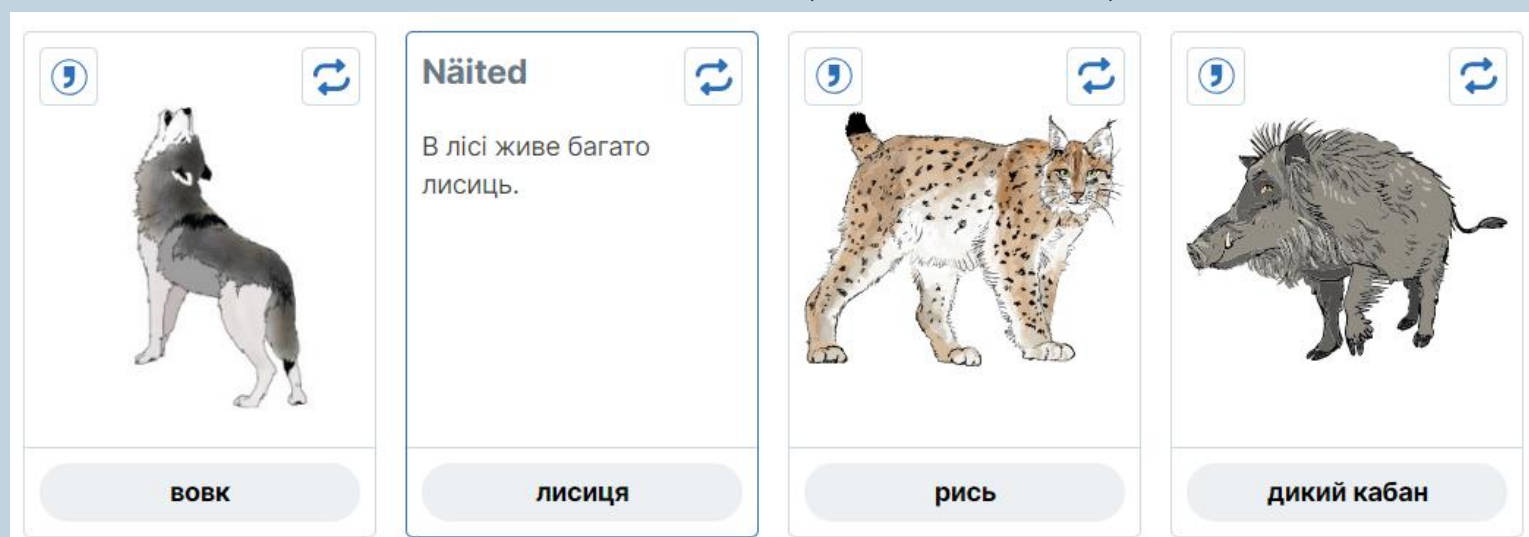
Text evaluation in Teacher Tools
sonaveeb.ee/teacher-tools

Picture Dictionary

1000 pictures

50 topics, e.g. animals, birds, furniture, food, clothing, bugs, health, vehicles, music instruments, body parts, etc.

Can be used in Estonian, Russian, Ukrainian



<https://sonaveeb.ee/wordgame>



References

Jürviste, Madis; Kallas, Jelena; Langemets, Margit; Tuulik, Maria; Viks, Ülle (2011). **Extending the functions of the EELEX dictionary writing system using the example of the Basic Estonian Dictionary**. Electronic Lexicography in the 21st Century New Applications for New Users: Proceedings of eLex 2011, Bled, 10-12 November 2011. Ed. Kosem; Iztok; Kosem, Karmen. Ljubljana: Trojina, Institute for Applied Slovenian Studies, 106–112.

Koppel, Kristina; Tavast, Arvi; Langemets, Margit; Kallas, Jelena (2019). **Aggregating dictionaries into the language portal Sõnaveeb: issues with and without a solution**. In: Kosem, I., Zingano Kuhn, T., Correia, M., Ferreria, J. P., Jansen, M., Pereira, I., Kallas, J., Jakubíček, M., Krek, S. & Tiberius, C. (Ed.). Proceedings of the eLex 2019 conference. 1-3 October 2019, Sintra, Portugal.. (434–452). Brno: Lexical Computing CZ, s.r.o.

Langemets, Margit; Loopmann, Andres; Viks, Ülle (2010). **Dictionary management system for bilingual dictionaries**. In: Sylviane Granger, Magali Paquot (Ed.). eLexicography in the 21st century : New challenges, new applications. Proceedings of eLex 2009, Louvain-la-Neuve, 22-24 October 2009. (425–430). Louvain-la-Neuve: Presses universitaires de Louvain, Cahiers du CENTAL.

Tavast, Arvi; Langemets, Margit; Kallas, Jelena; Koppel, Kristina (2018). **Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX**. Proceedings of the XVIII EURALEX International Congress: EURALEX: Lexicography in Global Contexts, Ljubljana, 17-21 July 2018. Ed. Jaka Čibej, Vojko Gorjanc, Iztok Kosem, Simon Krek. Ljubljana University Press, Faculty of Arts, 749–761.

Thank you