EESTI KEELE INSTITUUT

From online dictionaries to language infrastructure: developments in Estonia

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

20th EFNIL Conference
12 October 2023

EESTI KEELE INSTITUUT

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**

EESTI KEELE INSTITUUT

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

EESTI KEELE INSTITUUT

Dictionaries 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

EESTI KEELE INSTITUUT

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**

EESTI
KEELE
INSTITUUT

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.

EESTI KEELE INSTITUUT

Leksikograafid | Uus keelend | Lihtotsing | Detailotsing | Kristina Koppel, EKI ühendsõnastik 2023 (...

Valitud sõnakogusid: 124 | auto | Otsi | Otselink

Kokku leiti: 2

1 auto (et)
eki, les, mil
Avalik
süno valmis

2 auto- (et)
õtb

auto | Ühenda | Märgi tehtuks | 28.07.2023 | 28.07.2023 | Logi

eesti | + Vormikood | + Hääldus | 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]

EKI ühendsõnastik 2023

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

ÕS 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	SgIll - ainsuse sisseütlev
auto[s	SgIn - ainsuse seesütlev
auto[st	SgEl - ainsuse seestütlev
auto[le	SgAll - ainsuse alaleütlev
auto[l	SgAd - ainsuse alalütlev
auto[lt	SgAbl - 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	PlN - mitmuse nimetav
auto[de	PlG - mitmuse omastav
auto[sid	PlP - mitmuse osastav
auto[desse	Plill - mitmuse sisseütlev
auto[des	PlIn - mitmuse seesütlev
auto[dest	PlEl - mitmuse seestütlev
auto[dele	PlAll - mitmuse alaleütlev
auto[del	PlAd - mitmuse alalütlev
auto[delt	PlAbl - mitmuse alaltütlev
auto[deks	PlTr - mitmuse saav
auto[deni	PlTer - mitmuse rajav
auto[dena	PlEs - mitmuse olev
auto[deta	PlAb - mitmuse ilmaütlev
auto[dega	PlKom - mitmuse kaasaütlev

Sõna seos
Ühendid
Veel sarnaseid sõnu

www.ekilex.ee

EESTI KEELE INSTITUUT

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

EESTI KEELE INSTITUUT

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.)

EESTI KEELE INSTITUUT

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

EESTI
KEELE
INSTITUUT

Sõnaveeb

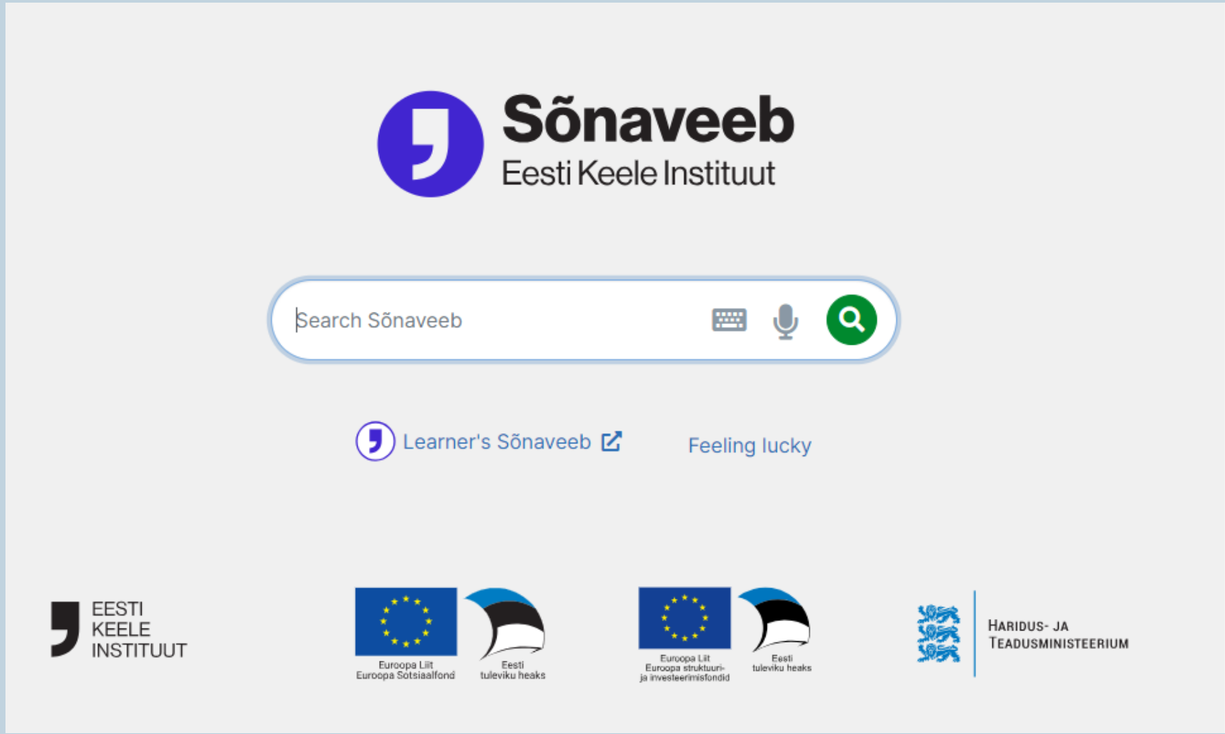
Language portal Sõnaveeb

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



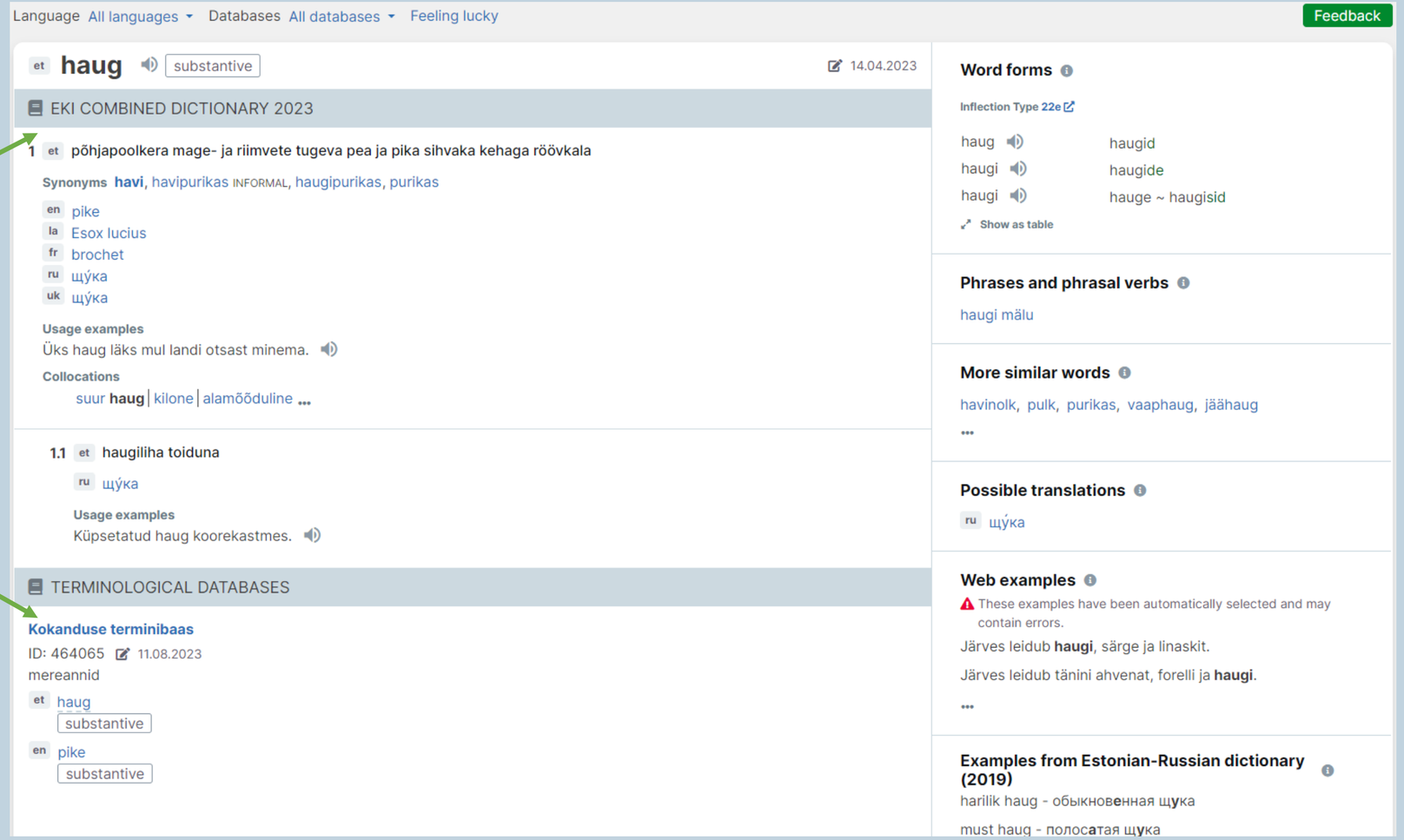
www.sonaveeb.ee

EESTI KEELE INSTITUUT

Sõnaveeb

Primarily for native speakers

Displays both general and specialized language



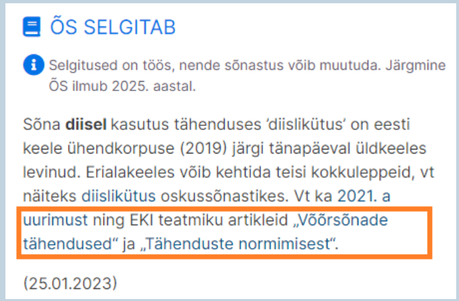
EKI Reference Book (EKI teatmik)

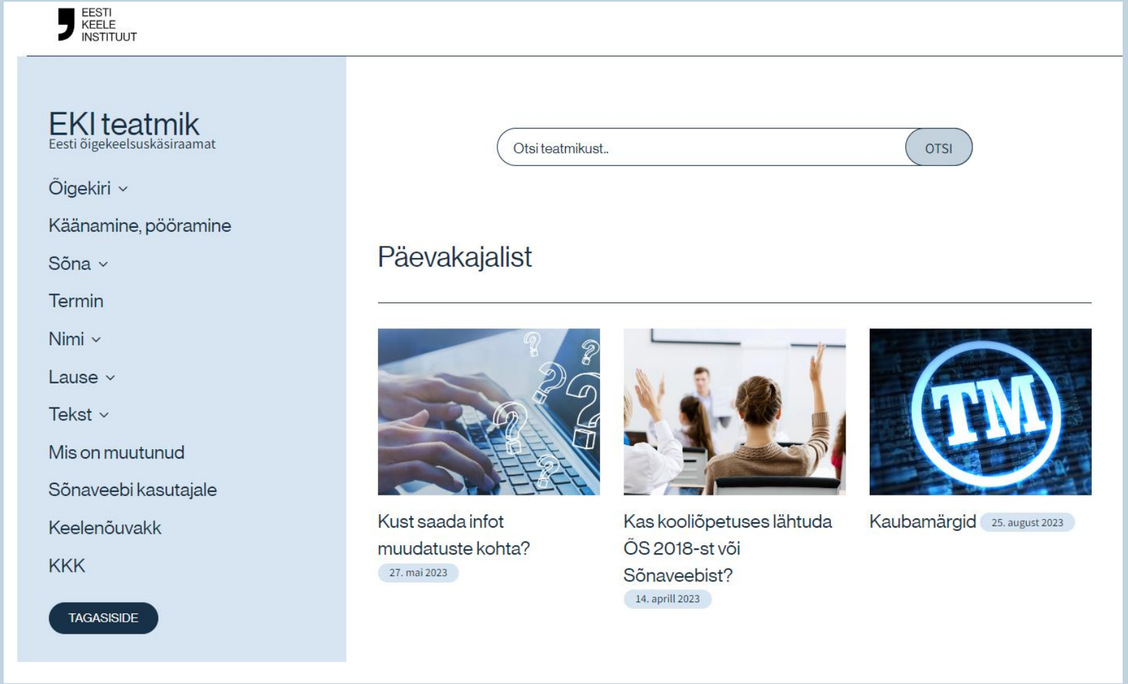
A collection of texts serving for information and advice (e.g. the approved orthographic rules)

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

EESTI KEELE INSTITUUT

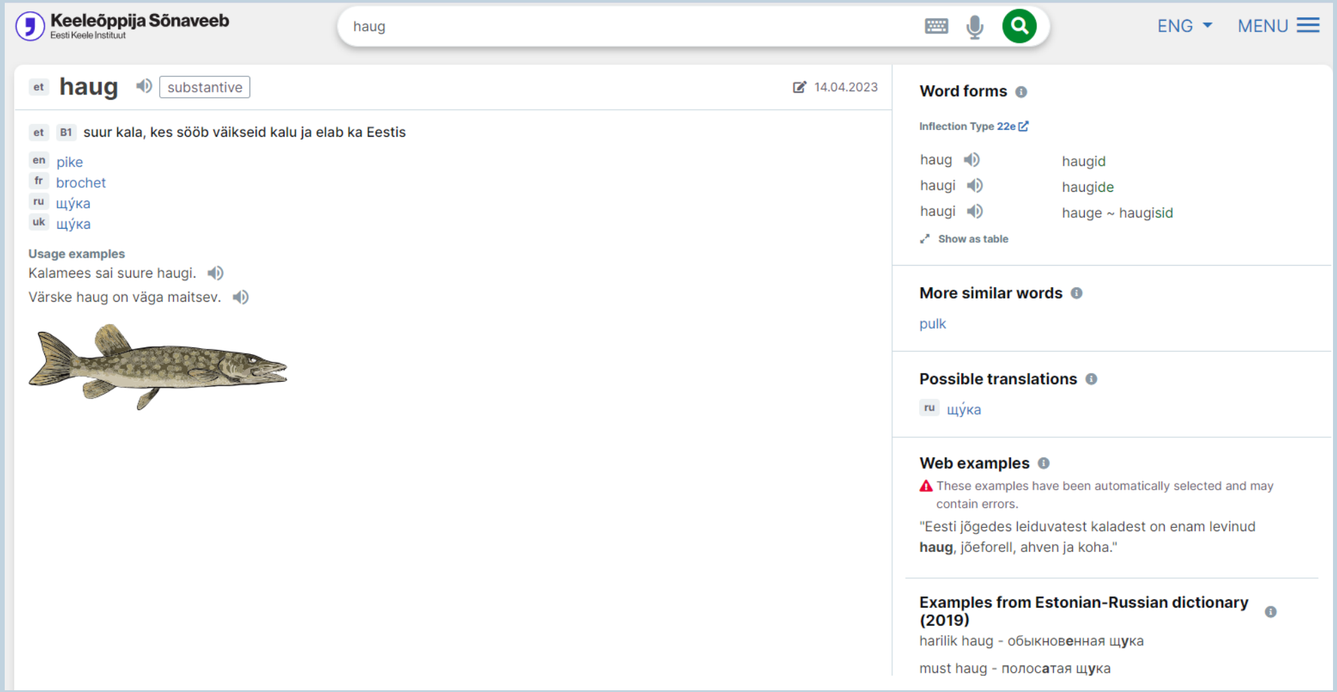
Tools for learners and teachers of Estonian L2

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



www.sonaveeb.ee/lite

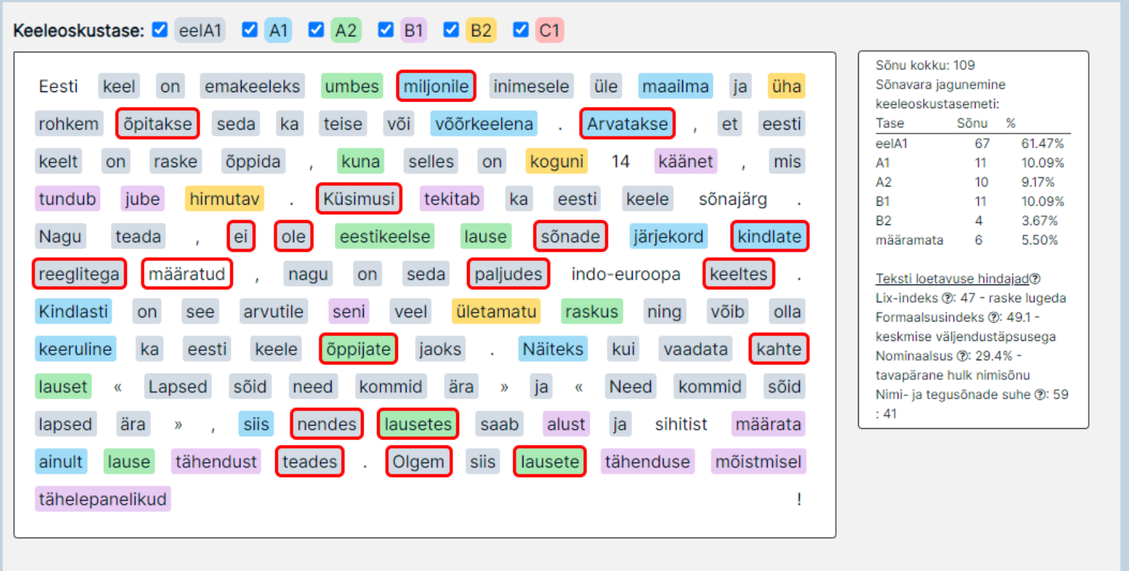
EESTI KEELE INSTITUUT

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



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

EESTI KEELE INSTITUUT

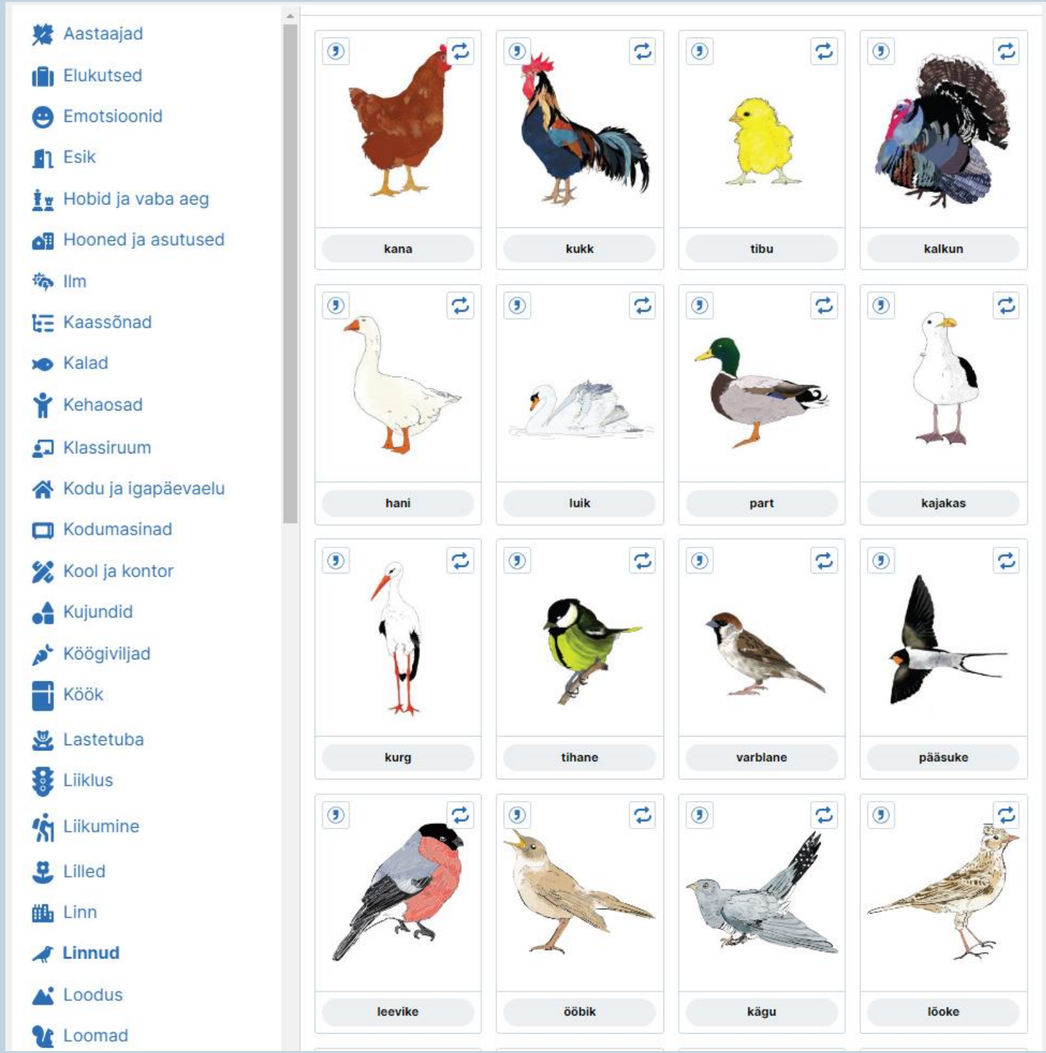
Picture Dictionary

1000 pictures

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

Can be used in Estonian, Russian, Ukranian

https://sonaveeb.ee/wordgame



EESTI KEELE INSTITUUT

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.

EESTI
KEELE
INSTITUUT

Thank you