CLARIN in Estonia – closer to Digital Humanities or Language Technology?

Kadri Vider
Kadri.vider@ut.ee
Center of Estonian Language Resources
University of Tartu, Estonia

CELR

Center of Estonian Language Resources (CELR, ee.clarin.eu)

- The research infrastructure of national importance in the Estonian Research Infrastructures Roadmap
  - To **make Estonian language data and tools available** to everyone working with digital language materials
  - To coordinate and organise the **documentation** and **archiving** of the language data and tools
  - To develop and promote LR **standards in Estonia**
  - To draw up necessary legal contracts and **licences** for **different types of users**: public, academic, commercial, etc
- **CELR performs the obligations of Estonia as the member in CLARIN ERIC**

ee.clarin.eu

CELR

Euroopa Liit
Euroopa
Regionaalarengu Fond

Eesti tuleviku heaks

Center of Estonian Language Resources (CELR)

- Consortium of 4 partners: **University of Tartu, TalTech, Institute of Estonian Language** and **Estonian Literary Museum**
- Archives and makes available LR and LT, incl. all outcomes of NPs “Estonian Language Technology” and “Estonian Language and Culture in the Digital Age”
- Catalogue and META-SHARE Registry of Estonian LRs and LTs
- Provides know-how and expertise about Estonian LR and LT, but also supports researchers with Data Management Planning, data and (web) tools in Digital Humanities

CELR

EESTI KEELETEHNOLOOGIA

NATIONAL PROGRAMME FOR ESTONIAN LANGUAGE TECHNOLOGY

"…to achieve a level of language technology support for the Estonian language to enable the language to successfully operate and thrive in today's information technology-based world. "

2006-2010: 33 R&D projects for software prototypes and language resources

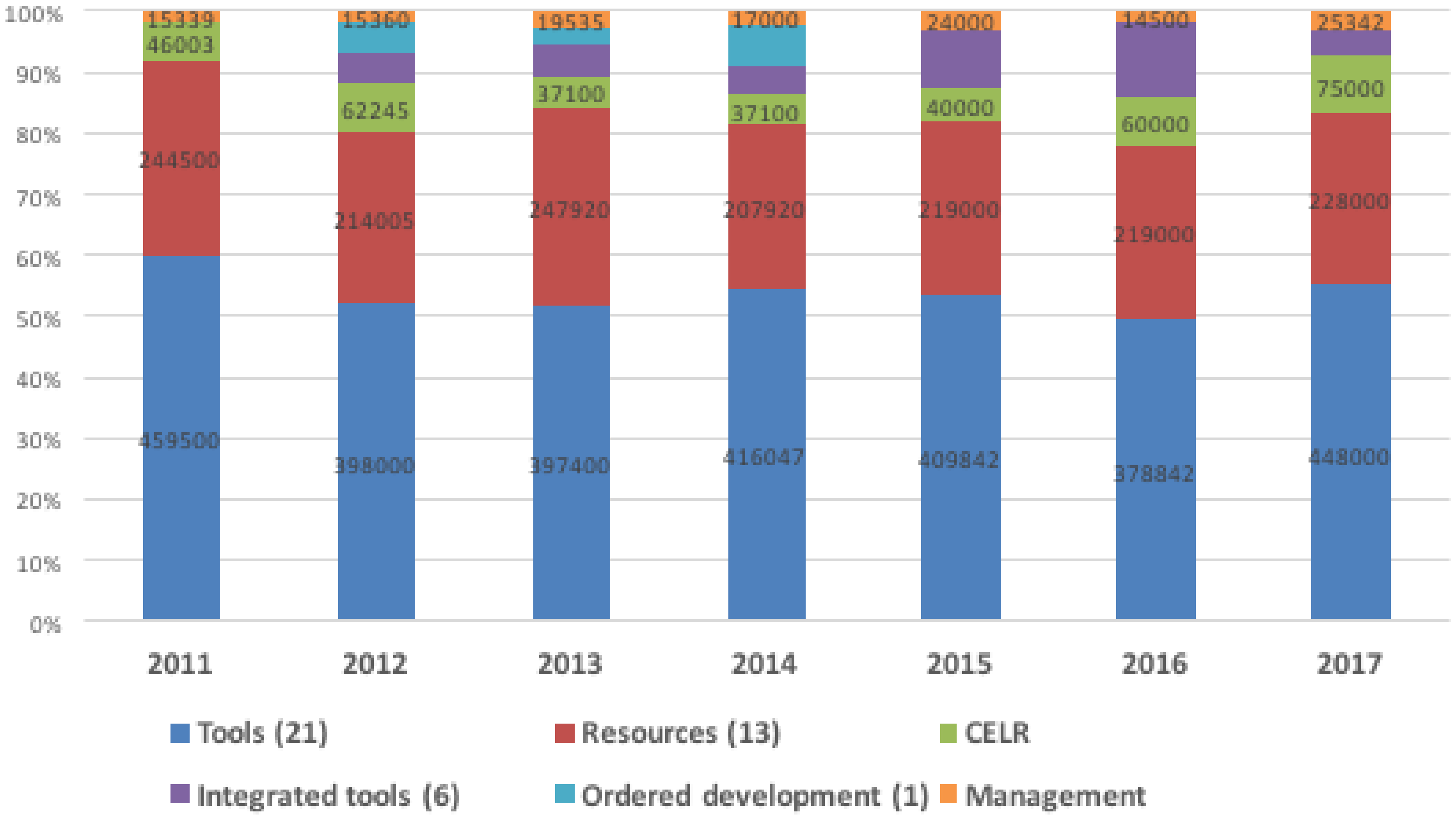
2011-2017: 43 R&D projects in 5 sub-objectives

+ **CELR** has an obligation to manage and to deposit all resources and tools developed within NPELT for preservation and long-term access

All project results (software and resources) are **Open Access, Open Data and Open Source as much as possible**.

www.keeletehnoloogia.ee

NPELT (2011 - 2017) 765 342 €/year



EESTI KEELETEHNOLOOGIA
NATIONAL PROGRAMME FOR ESTONIAN
LANGUAGE TECHNOLOGY

Current period 2018-2027

1) basic technologies and resources:

- Estonian Open Parallel Corpora
- Phonetic Corpus of Spontaneous Estonian Speech III
- Estonian Universal Syntax: Resources and Applications
- Estonian Wordnet as applicable language resource
- Further development of EstNLTK library and related web services
- Speech recognition
- Grammar checker prototype
- Neurotõlge: Adaptive, Multilingual and Reliable Machine Translation for Estonian
- Neural network based text analysis models for Estonian
- Diversification of Estonian text-to-speech synthesis

2) applications of technology:

- TEXTA Toolkit 2.0
- Võro and Seto integrated language technology and language resources

CELR

www.keeletehnoloogia.ee

Estonian Language and Culture in the Digital Age (2019–2027)

- compiling monolingual and bilingual dictionaries and comprehensive studies
- Support the use of language technology tools in studying the language.
- Support the prevalence of language technology applications
- Support wider implementation of data collections and technologies

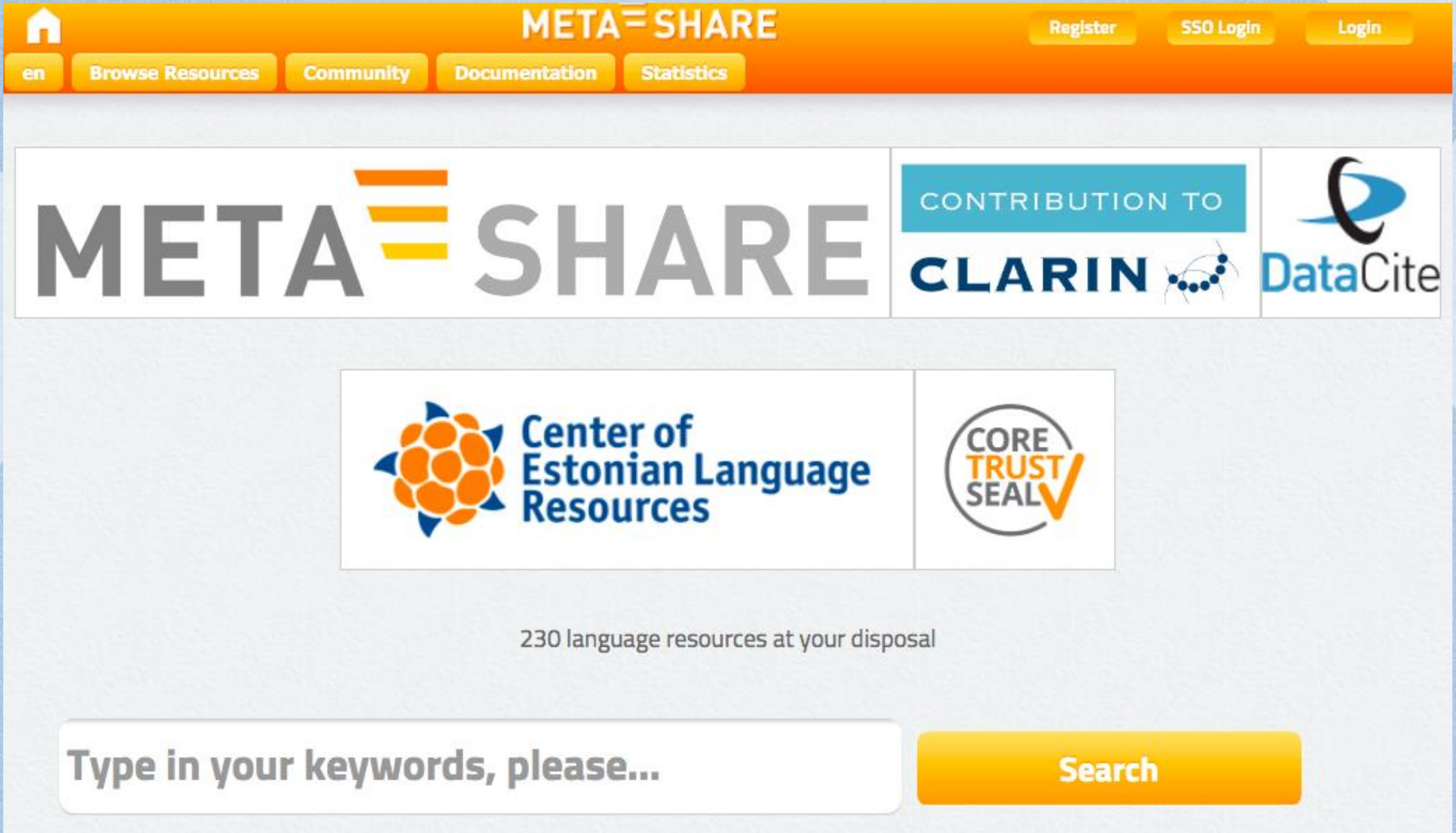
From subobjectives of the programme:
2. Introducing **digital tools in research**, especially results of NPELT
4. The results being **as widely accessible as possible** - publishing research data in repositories, and introducing their (re-)use as **FAIR Open Data**.

Estonian Language and Culture in the Digital Age 2019–2027

Current projects

- Lexis and planning of Estonian: descriptive and prescriptive aspects – Institute of Estonian Language
- Source documents in the cultural process: Estonian materials in the collections and databases of the Estonian Literary Museum - Estonian Literary Museum
- Data and corpora of Estonian children and youth multilingual communication – University of Tallinn
- Possibilities of automatic analysis of historical texts by the example of 19th-century Estonian communal court minutes - UT
- The Ethnic History of Estonian Peoples in the light of new research
- Digi-OWLDI: Five centuries of written Estonian vocablulary, morphology and phonology
- Interdisciplinary corpus of Seto
- Teen speak in Estonia
- The prosody and information structure of surprise questions in Estonian in comparison with other languages

Registry for LR-s - META-SHARE

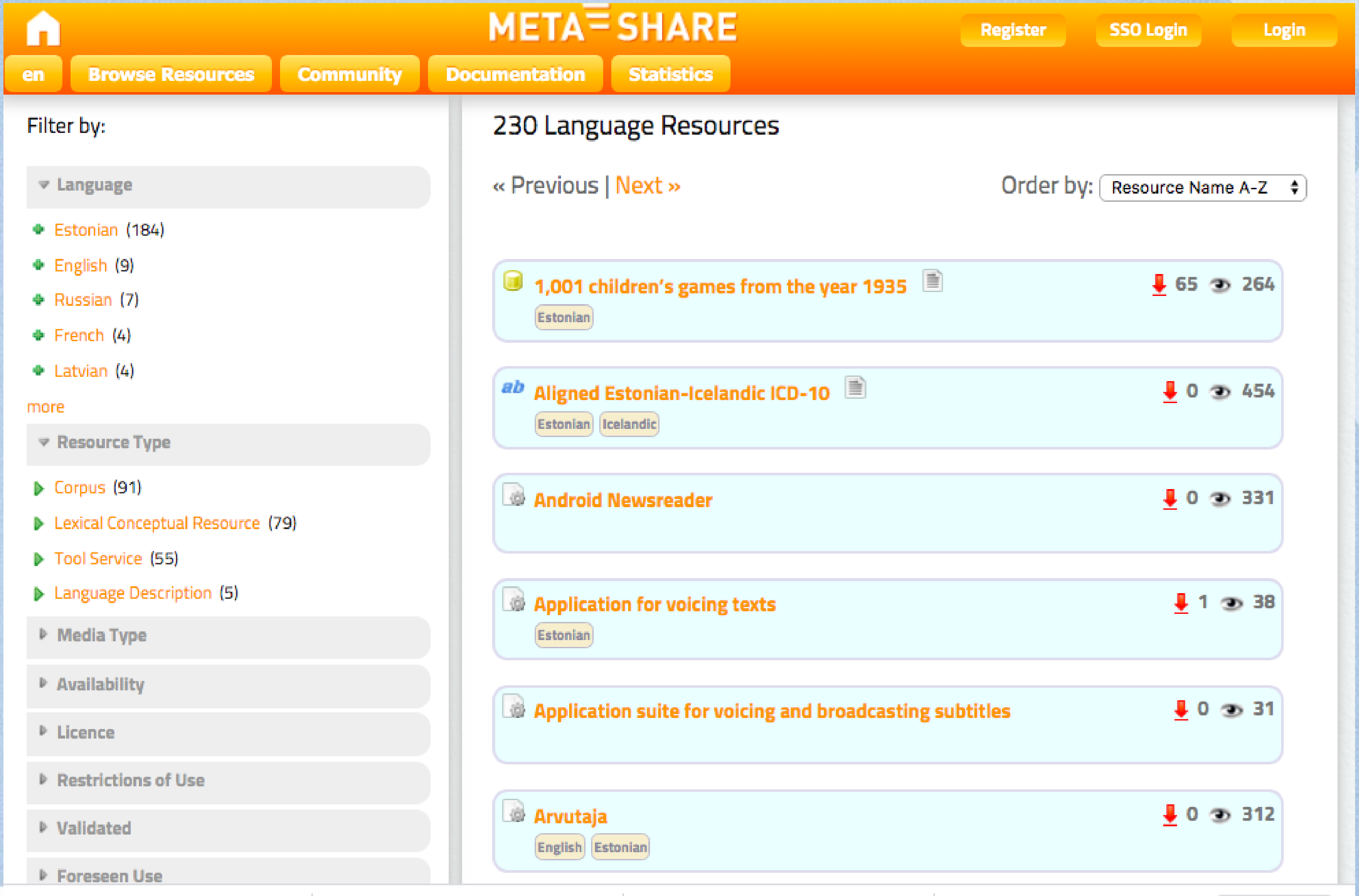


230 language resources at your disposal

CELR

https://metashare.ut.ee

Registry for LR-s - META-SHARE



Filter by:

▾ Language

- Estonian (184)
- English (9)
- Russian (7)
- French (4)
- Latvian (4)

more

▾ Resource Type

- Corpus (91)
- Lexical Conceptual Resource (79)
- Tool Service (55)
- Language Description (5)

▸ Media Type

▸ Availability

▸ Licence

▸ Restrictions of Use

▸ Validated

▸ Foreseen Use

230 Language Resources

« Previous | Next »

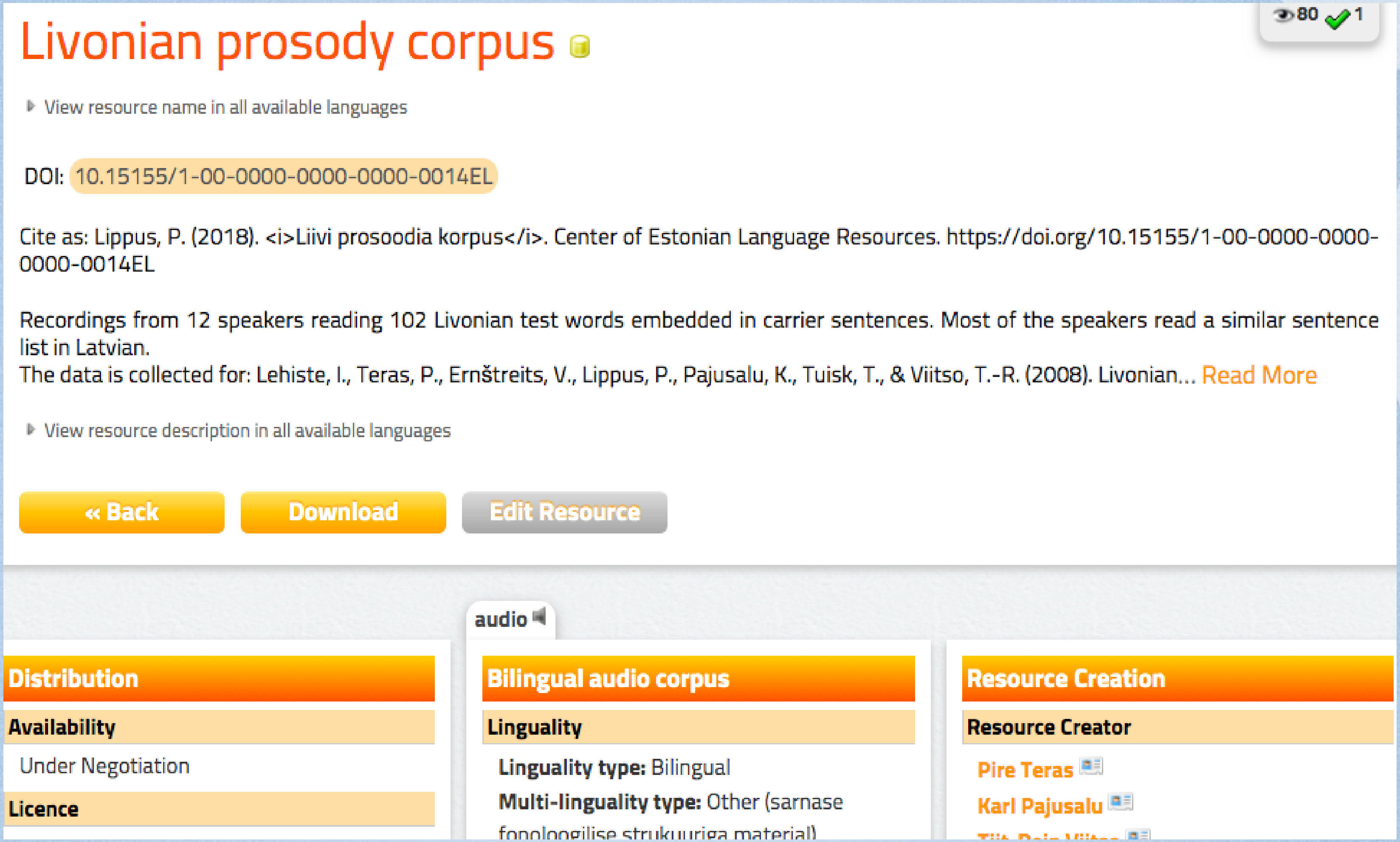
Order by: Resource Name A-Z

Resource	Languages	Downloads	Views
1,001 children's games from the year 1935	Estonian	65	264
Aligned Estonian-Icelandic ICD-10	Estonian, Icelandic	0	454
Android Newsreader		0	331
Application for voicing texts	Estonian	1	38
Application suite for voicing and broadcasting subtitles		0	31
Arvutaja	English, Estonian	0	312

CELR

https://metashare.ut.ee

Registry for LR-s - META-SHARE



👁80 ✓1

Livonian prosody corpus

▸ View resource name in all available languages

DOI: 10.15155/1-00-0000-0000-0000-0014EL

Cite as: Lippus, P. (2018). <i>Liivi prosoodia korpus</i>. Center of Estonian Language Resources. https://doi.org/10.15155/1-00-0000-0000-0000-0014EL

Recordings from 12 speakers reading 102 Livonian test words embedded in carrier sentences. Most of the speakers read a similar sentence list in Latvian.
The data is collected for: Lehiste, I., Teras, P., Ernštreits, V., Lippus, P., Pajusalu, K., Tuisk, T., & Viitso, T.-R. (2008). Livonian... Read More

▸ View resource description in all available languages

Distribution

Availability

Under Negotiation

Licence

Bilingual audio corpus

Linguality

Linguality type: Bilingual

Multi-linguality type: Other (sarnase fonoloogilise strukuuriga material)

Resource Creation

Resource Creator

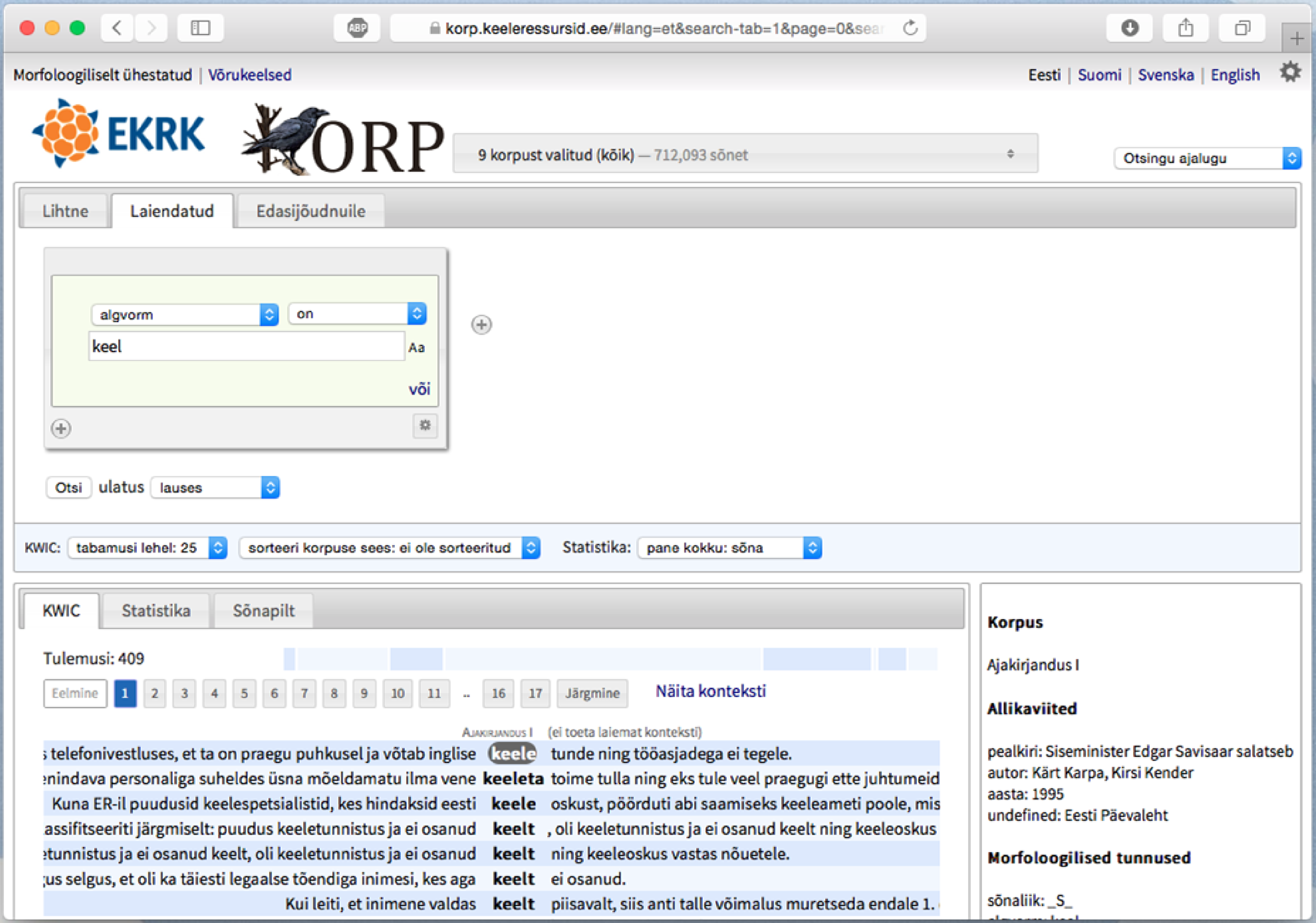
Pire Teras

Karl Pajusalu

Tiit-Rein Viitso

https://metashare.ut.ee

Corpus Query Service KORP



9 korpust valitud (kõik) — 712,093 sõnet

algvorm on keel

ulatus: lauses

KWIC: tabamusi lehel: 25 sorteeri korpuse sees: ei ole sorteeritud Statistika: pane kokku: sõna

Tulemusi: 409

AJAKIRJANDUS I (ei toeta laiemat konteksti)

; telefonivestluses, et ta on praegu puhkusel ja võtab inglise **keele** tunde ning tööasjadega ei tegele.
ɛnindava personaliga suheldes üsna mõeldamatu ilma vene **keeleta** toime tulla ning eks tule veel praegugi ette juhtumeid
Kuna ER-il puudusid keelespetsialistid, kes hindaksid eesti **keele** oskust, pöörduti abi saamiseks keeleameti poole, mis
assifitseeriti järgmiselt: puudus keeletunnistus ja ei osanud **keelt** , oli keeletunnistus ja ei osanud keelt ning keeleoskus
ɛtunnistus ja ei osanud keelt, oli keeletunnistus ja ei osanud **keelt** ning keeleoskus vastas nõuetele.
;us selgus, et oli ka täiesti legaalse tõendiga inimesi, kes aga **keelt** ei osanud.
Kui leiti, et inimene valdas **keelt** piisavalt, siis anti talle võimalus muretseda endale 1.

Korpus

Ajakirjandus I

Allikaviited

pealkiri: Siseminister Edgar Savisaar salatseb
autor: Kärt Karpa, Kirsi Kender
aasta: 1995
undefined: Eesti Päevaleht

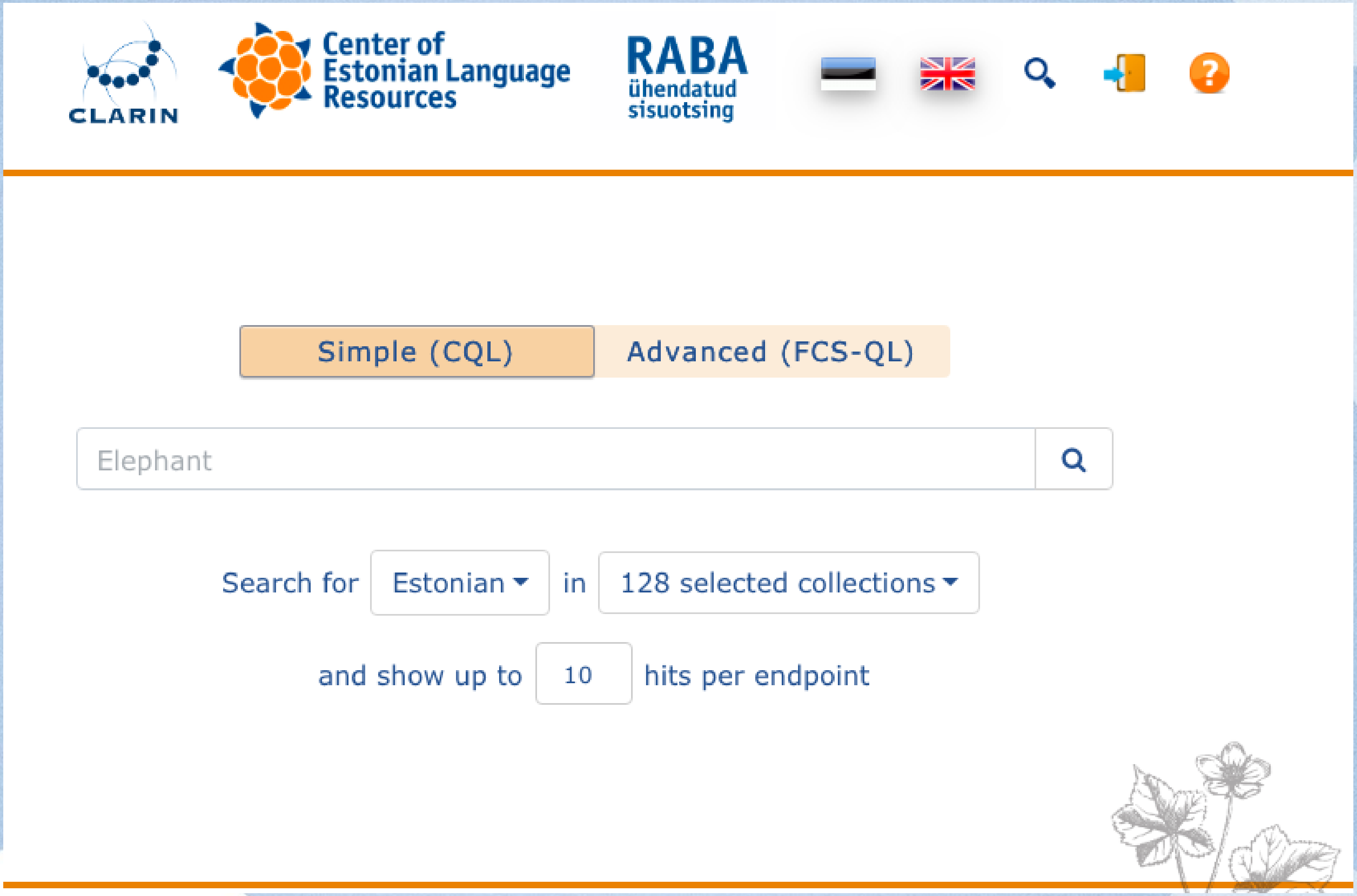
Morfoloogilised tunnused

sõnaliik: _S_

CELR

https://korp.keeleressursid.ee

Federated Content Search Raba



Elephant

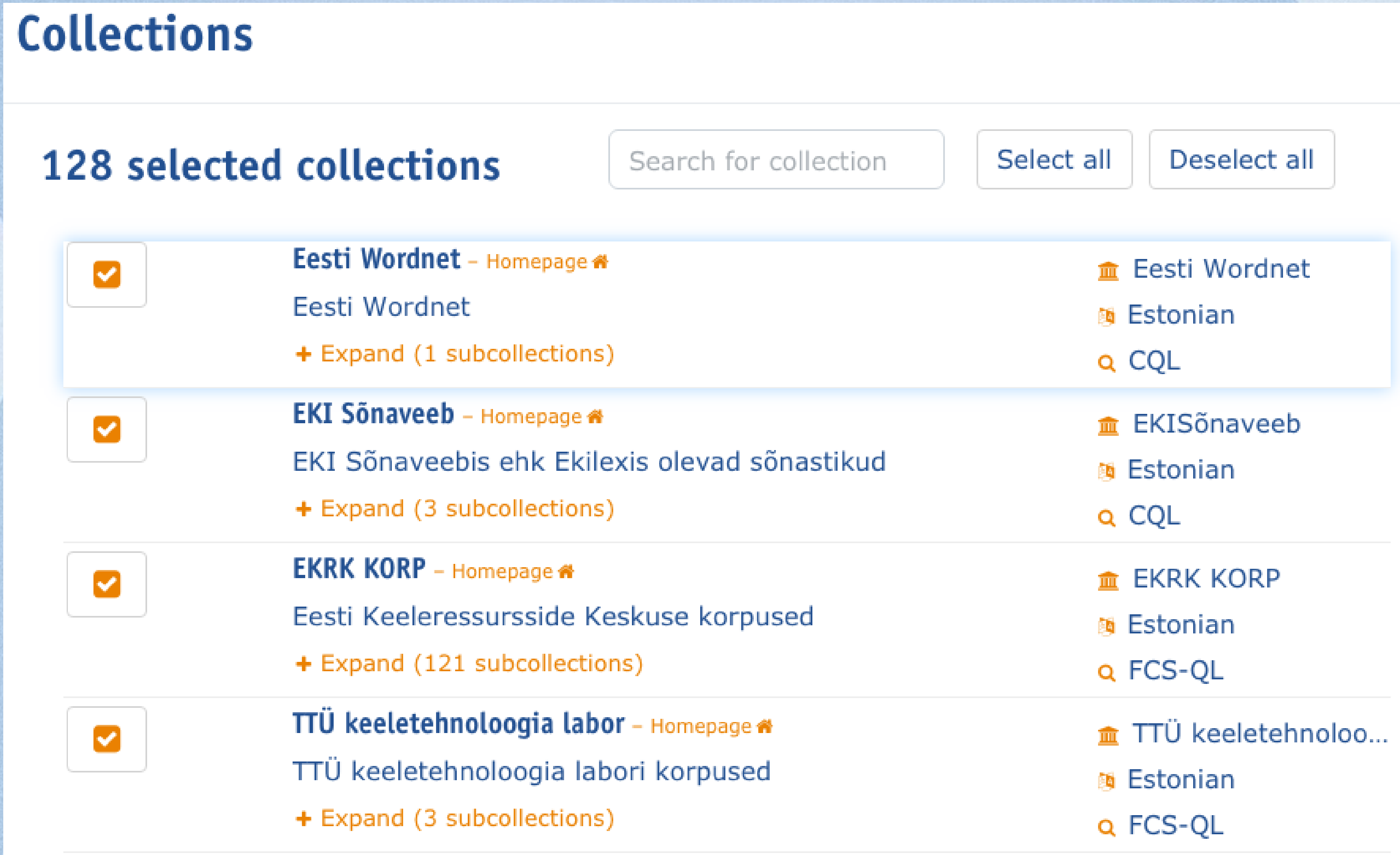
Search for Estonian in 128 selected collections

and show up to 10 hits per endpoint

CELR

https://raba.keeleressursid.ee/

Federated Content Search Raba



Collections

128 selected collections

☑	**Eesti Wordnet** – Homepage Eesti Wordnet + Expand (1 subcollections)	Eesti Wordnet Estonian CQL
☑	**EKI Sõnaveeb** – Homepage EKI Sõnaveebis ehk Ekilexis olevad sõnastikud + Expand (3 subcollections)	EKISõnaveeb Estonian CQL
☑	**EKRK KORP** – Homepage Eesti Keeleressursside Keskuse korpused + Expand (121 subcollections)	EKRK KORP Estonian FCS-QL
☑	**TTÜ keeletehnoloogia labor** – Homepage TTÜ keeletehnoloogia labori korpused + Expand (3 subcollections)	TTÜ keeletehnoloo... Estonian FCS-QL

CELR

https://raba.keeleressursid.ee/

Thank you!

www.keeletehnoloogia.ee

ee.clarin.eu

CELR

Lessons and risks of NPELT

- R&D projects within an open competition do not cover the full spectrum of goals to support Estonian language in technology
- Researchers are mostly interested in a result (prototype) rather than the stable application which can be integrated into software products
- Relation to IT business and production is weak
- NP does not deal explicitly with the education of new generation of language technologists
- Language data is often subject to copyright protection or sensitive personal data