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
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
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
“...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

- Tools (21)
- Resources (13)
- CELR
- Integrated tools (6)
- Ordered development (1)
- Management
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

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.

[CELR Logo]
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 vocabulary, 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

META-SHARE

CONTRIBUTION TO
CLARIN

DataCite

Center of Estonian Language Resources

CORE TRUST SEAL

230 language resources at your disposal

Type in your keywords, please...

Search

https://metashare.ut.ee
### Registry for LR-s - META-SHARE

230 Language Resources

- **1,001 children's games from the year 1935**
  - Estonian
  - Download: 65, Views: 264

- **Aligned Estonian-Icelandic ICD-10**
  - Estonian, Icelandic
  - Download: 0, Views: 454

- **Android Newsreader**
  - Download: 0, Views: 331

- **Application for voicing texts**
  - Estonian
  - Download: 1, Views: 38

- **Application suite for voicing and broadcasting subtitles**
  - Download: 0, Views: 31

- **Arvutaja**
  - English, Estonian
  - Download: 0, Views: 312

---

**Filter by:**

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

- **Resource Type**
  - Corpus (91), Lexical Conceptual Resource (79), Tool Service (55)

- **Media Type**

- **Availability**

- **Licence**

- **Restrictions of Use**

- **Valued**

- **Forseen Use**

**Order by:** Resource Name A-Z
Livonian prosody corpus

View resource name in all available languages

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

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

Recordings from 12 speakers reading 102 Livonian test words embedded in carrier sentences. Most of the speakers read a similar sentence list in Latvian.

View resource description in all available languages

Distribution
- Availability: Under Negotiation

Bilingual audio corpus
- Linguality
  - Linguality type: Bilingual
  - Multi-linguality type: Other (sarnase fonooloogilise struktuuringa material)

Resource Creation
- Resource Creator
  - Pire Teras
  - Karl Pajusalu
  - Tiit Paal-Viitso

https://metashare.ut.ee
Corpus Query Service KORP

https://korp.keeleressursid.ee
# Federated Content Search Raba

## Collections

<table>
<thead>
<tr>
<th>128 selected collections</th>
<th>Search for collection</th>
<th>Select all</th>
<th>Deselect all</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Collection</th>
<th>Description</th>
<th>Selected</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Eesti Wordnet</strong> – Homepage</td>
<td>Eesti Wordnet</td>
<td>✔️</td>
<td>Expand (1 subcollections)</td>
</tr>
<tr>
<td><strong>EKI Sõnaveeb</strong> – Homepage</td>
<td>EKI Sõnaveebis ehk Ekilexis olevad sõnastikud</td>
<td>✔️</td>
<td>Expand (3 subcollections)</td>
</tr>
<tr>
<td><strong>EKRK Korp</strong> – Homepage</td>
<td>Eesti Keeleressursside Keskuse korpusid</td>
<td>✔️</td>
<td>Expand (121 subcollections)</td>
</tr>
<tr>
<td><strong>TTÜ keeleheloogia labor</strong> – Homepage</td>
<td>TTÜ keeleheloogia labori korpusid</td>
<td>✔️</td>
<td>Expand (3 subcollections)</td>
</tr>
</tbody>
</table>

[CELK](https://raba.keeleressursid.ee/)
Thank you!

www.keeletehnoloogia.ee

ee.clarin.eu
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