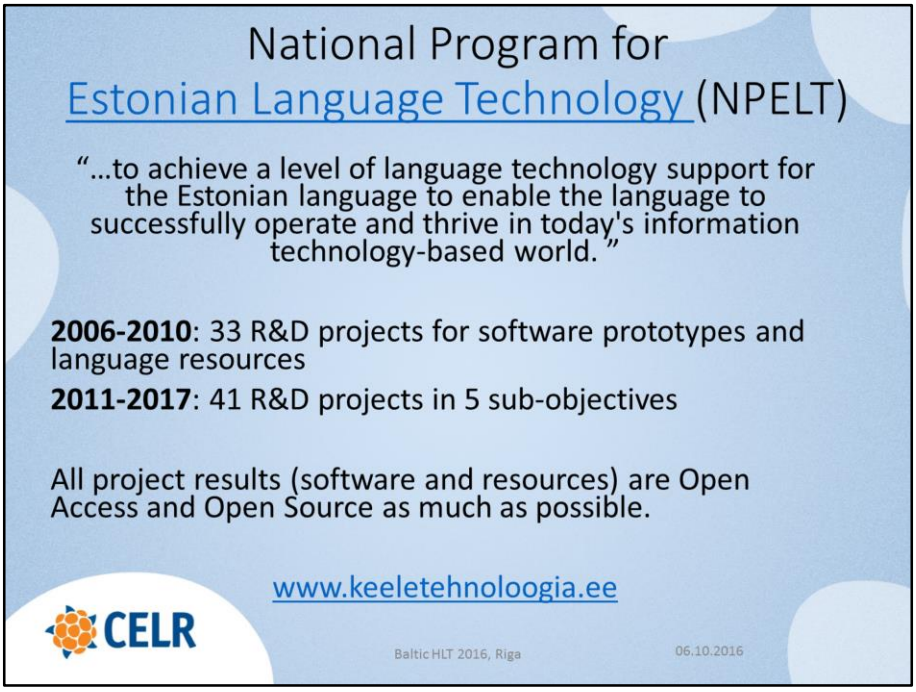
Estonian Language Resources and Tools: Overview

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



The National Programme for Estonian Language Technology (NPELT) funds language technology research and development, from the compilation of language data resources to the creation of software application prototypes. The focus is also on making the Estonian language digital resources (language resources and software) freely available.
NPELT aims 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.

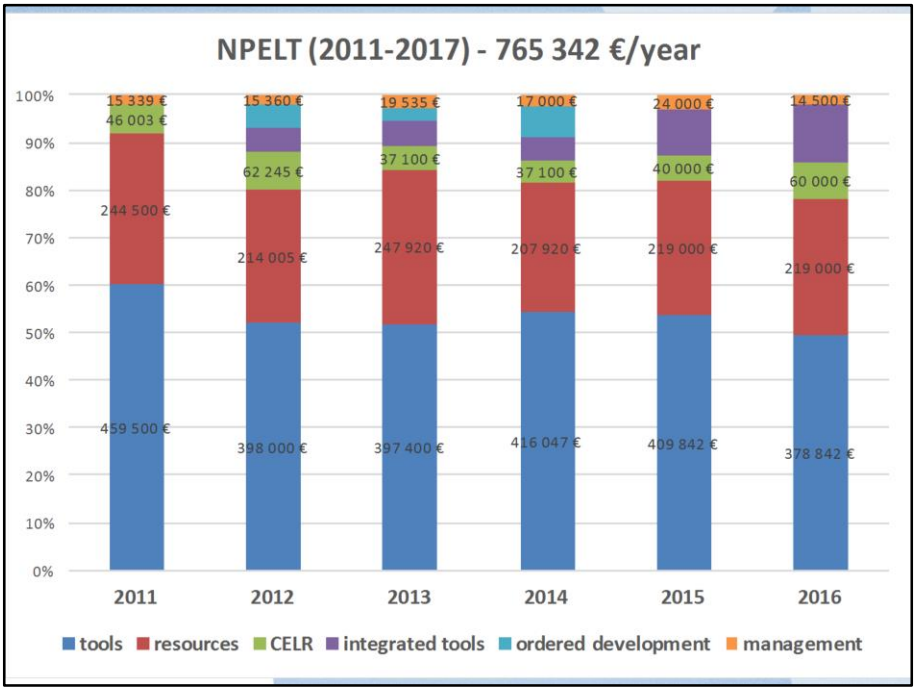
NPELT is divided into 5 sub-objectives (https://www.keeletehnoloogia.ee/en/projects-2011-2017)
Research and development projects for building software prototypes – the implementation of results is quite broad: from software prototypes to component software, not as often applications for end users. The most successful one is "Speech recognition" project building web and mobile applications in TTU Institute of Cybernetics. Worth mentioning is also Application Suite for voicing and broadcasting subtitles on television project in the Institute of Estonian Language.
Projects for building language resources – provides digital language data (text

corpora, dictionaries) for everyday language user and for language research as well as to IT-applications. These projects also relate to topics of big data and digital cultural heritage. Some remarkable results: Keeleveeb portal (www.keeleveeb.ee), which involves many different corpora and lexicons and uses lemmatization; also Estonian WordNet, (http://www.cl.ut.ee/ressursid/teksaurus/).

Center of Estonian Language Resources (CELR)
has an obligation to manage and to deposit all resources and tools developed within NPELT for preservation and long-term access (http://keeleressursid.ee/en/

Integrated language software and its applications - a prerequisite for the project application is the involvement of partners from the public or private sector. As the result for example technical aids for people with special needs and interface for public services are expected. Yet to be used more widely is a notable project in Institute of Estonian Language called „Generation of Audiobooks and voicing interface of Digar“.

Development projects to be ordered – Software development projects ordered on the proposal of the steering committee of NPELT. In 2013-2014 for example the open-source morphological analyzer software was ordered from the only one Estonian fully language technology company Filosoft (www.filosoft.ee).



NPELT is divided into 5 sub-objectives (https://www.keeletehnoloogia.ee/en/projects-2011-2017)

Research and development projects for building software prototypes – the implementation of results is quite broad: from software prototypes to component software, not as often applications for end users. The most successful one is "Speech recognition" project building web and mobile applications in TTU Institute of Cybernetics. Worth mentioning is also Application Suite for voicing and broadcasting subtitles on television project in the Institute of Estonian Language.

Projects for building language resources – provides digital language data (text corpora, dictionaries) for everyday language user and for language research as well as to IT-applications. These projects also relate to topics of big data and digital cultural heritage. Some remarkable results: Keeleveeb portal (www.keeleveeb.ee), which involves many different corpora and lexicons and uses lemmatization; also Estonian WordNet, (http://www.cl.ut.ee/ressursid/teksaurus/).

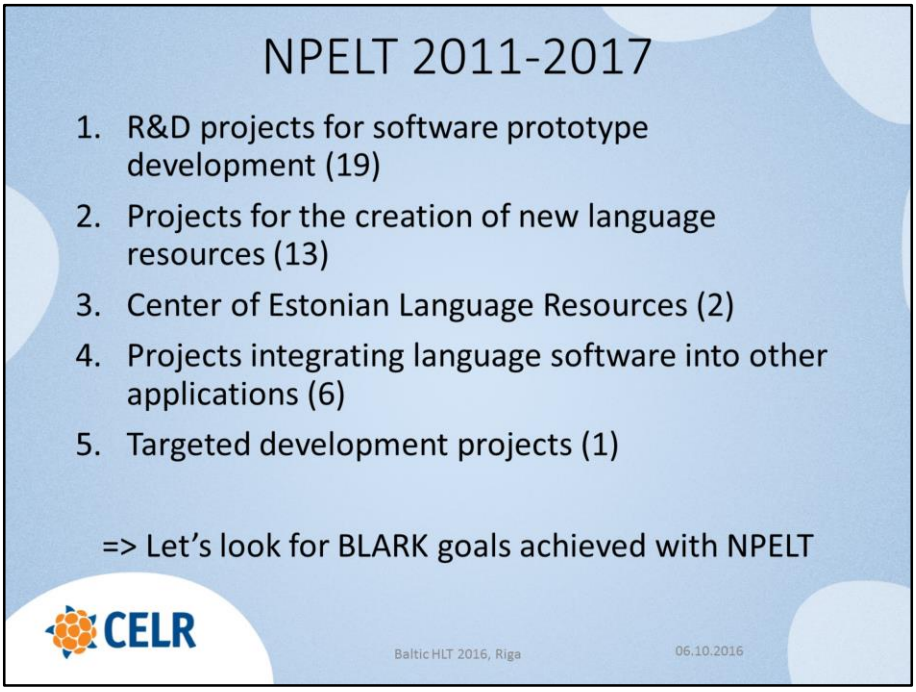
Center of Estonian Language Resources (CELR)

has an obligation to manage and to deposit all resources and tools developed within NPELT for preservation and long-term access (http://keeleressursid.ee/en/

Integrated language software and its applications - a prerequisite for the project

application is the involvement of partners from the public or private sector. As the result for example technical aids for people with special needs and interface for public services are expected. Yet to be used more widely is a notable project in Institute of Estonian Language called „Generation of Audiobooks and voicing interface of Digar“.

Development projects to be ordered – Software development projects ordered on the proposal of the steering committee of NPELT. In 2013-2014 for example the open-source morphological analyzer software was ordered from the only one Estonian fully language technology company Filosoft (www.filosoft.ee).



The National Programme for Estonian Language Technology (NPELT) funds language technology research and development, from the compilation of language data resources to the creation of software application prototypes. The focus is also on making the Estonian language digital resources (language resources and software) freely available.
NPELT aims 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.

NPELT is divided into 5 sub-objectives (https://www.keeletehnoloogia.ee/en/projects-2011-2017)
Research and development projects for building software prototypes – the implementation of results is quite broad: from software prototypes to component software, not as often applications for end users. The most successful one is "Speech recognition" project building web and mobile applications in TTU Institute of Cybernetics. Worth mentioning is also Application Suite for voicing and broadcasting subtitles on television project in the Institute of Estonian Language.
Projects for building language resources – provides digital language data (text

corpora, dictionaries) for everyday language user and for language research as well as to IT-applications. These projects also relate to topics of big data and digital cultural heritage. Some remarkable results: Keeleveeb portal (www.keeleveeb.ee), which involves many different corpora and lexicons and uses lemmatization; also Estonian WordNet, (http://www.cl.ut.ee/ressursid/teksaurus/).

Center of Estonian Language Resources (CELR)
has an obligation to manage and to deposit all resources and tools developed within NPELT for preservation and long-term access (http://keeleressursid.ee/en/

Integrated language software and its applications - a prerequisite for the project application is the involvement of partners from the public or private sector. As the result for example technical aids for people with special needs and interface for public services are expected. Yet to be used more widely is a notable project in Institute of Estonian Language called „Generation of Audiobooks and voicing interface of Digar“.

Development projects to be ordered – Software development projects ordered on the proposal of the steering committee of NPELT. In 2013-2014 for example the open-source morphological analyzer software was ordered from the only one Estonian fully language technology company Filosoft (www.filosoft.ee).

Speech technology

- ✓ Speech Recognition - Real time speech recognition, applications for smart phones and tablets, Estonian content search in audio (broadcasting) files - Laboratory of Phonetics and Speech Technology at TTU
- ✓ Speech Synthesis - Emotional speech synthesis, audio-visual speech synthesis i.e. 3D head – at Institute of Estonian Language and TTU
- – Real-time dialogue system with SR input and TtS output

CELR

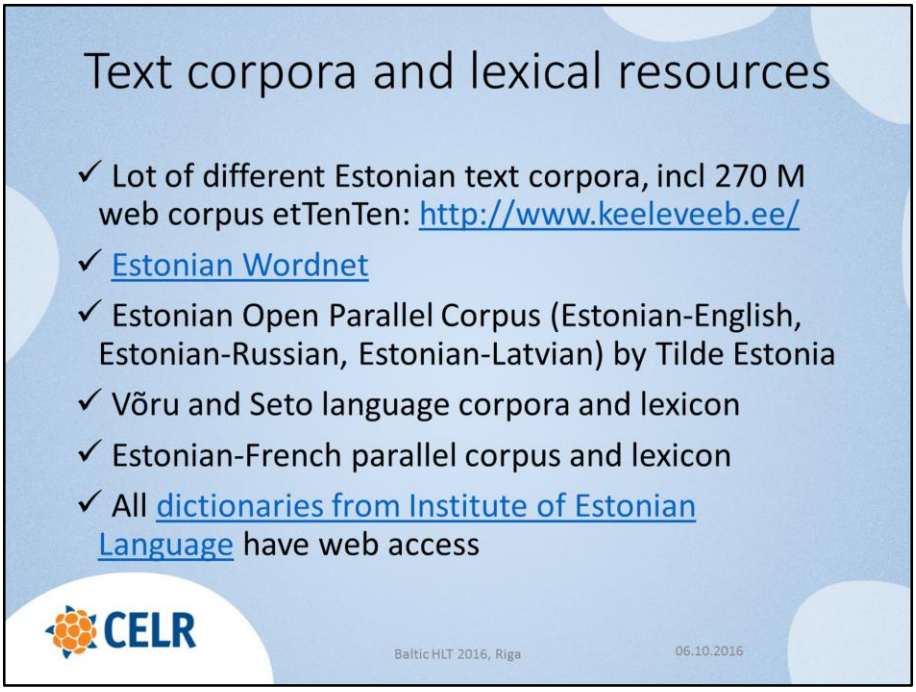
Machine translation

- ✓ Statistical machine translation
- ✓ Post-edited machine translation
- ✓ Translation models trained on parallel corpora
- − Still the most well-known is Google

Applications for text analysis

- ✓ Vabamorf: Open Source software to analyse and synthesise the Estonian morphology
- ✓ Dependency parser, collocation detector
- ✓ EstNLTK: Open source tools in Python for Estonian natural language processing
- ✓ Template-based mining of facts from a text corpus
- ✓ Detector of text polarity
- ✓ Framework for creating domain ontologies from text corpora (mainly in medicine)
- – Grammar corrector

CELR



etTenTen

etTenTen korpus on internetist alla laetud eestikeelsete veebilehtede korpus. Korpuses on 270 miljonit sõna 686 000 veebilehelt.

Compilation and editing of dictionaries

- ✓ EELex: tool for compiling dictionaries on paper and on the Web
- ✓ Language hotline to support the system of e-dictionaries, language hotline and language planning

CELR

Just some examples

Integrated applications

- ✓ Application suite for **voicing** and broadcasting **subtitles** on television (in co-operation with the Estonian Public Broadcasting service and the Estonian Association for the Blind)
- ✓ Audiobook generator and Digar audiointerface (in co-operation with the National Library)

CELR

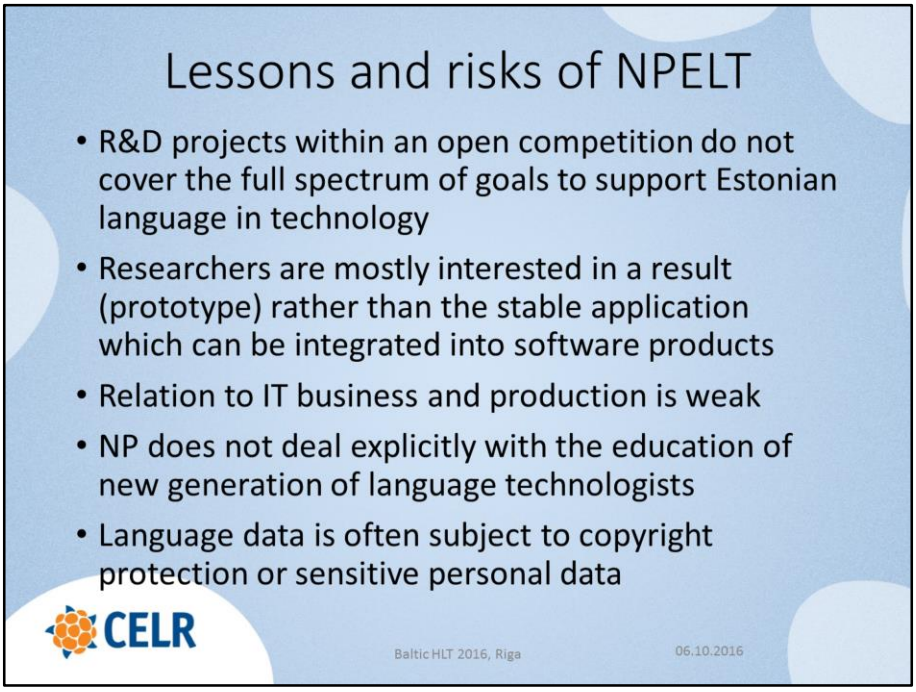
Dialogue Systems

- ✓ Estonian workbench for dialogue systems
- − automatic recognition/synthesis of dialogue acts
- − automatic recognition/synthesis of dialogue strategies

Analysis-synthesis of cohesive texts

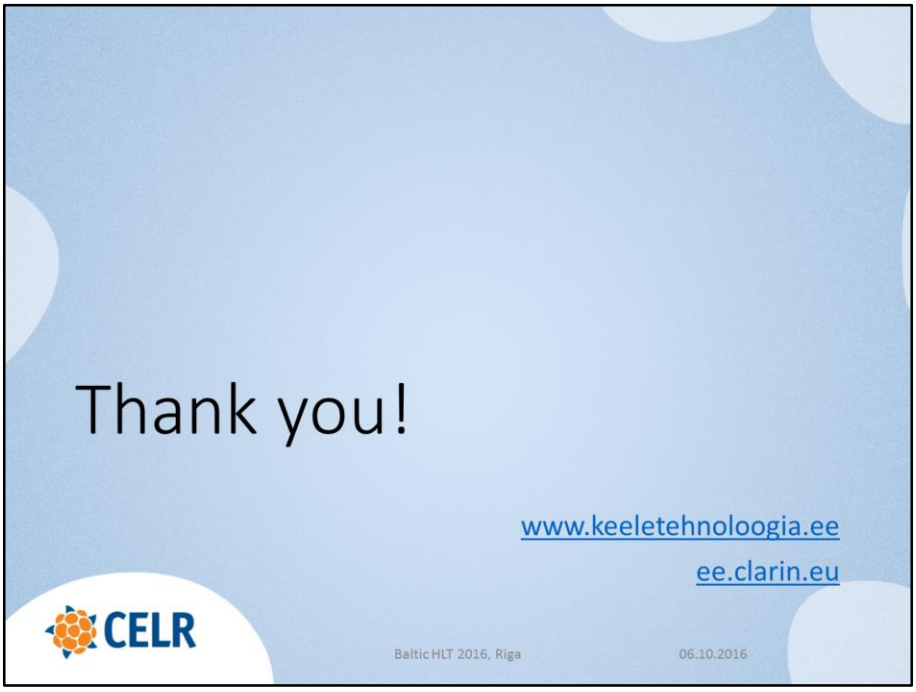
- ✓automatic recognition of extralinguistic communication signals (voice quality, emotions, laughter, etc.)
- ? (semi-)automatic transcription system for spoken language
- – automatic recognition/synthesis of cohesive text structure, means for coherence and categorisation in text, dialogue structure (separate for spoken and written, e.g. the Internet dialogue)
- – syntactic analysis-synthesis of cohesive texts
- – semantic analysis of compound sentences and cohesive texts (in specific fields) and means for semantic synthesis
- – pragmatic analysis-synthesis (in specific fields) and integration with other levels of language (syntax, semantics)

CELR



Risks in NPELT to achieve the goals:

* Projects are applied within an open competition – ideas which inspire researchers do not fully cover the goals; and the support of Estonian language technology is not systematically developed;
* Research and development projects – 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: how to improve this situation, how to implement prototypes which support Estonian language on behalf of information society?
* NPELT does not deal explicitly with the education of new generation of language technologists; we lack from the knowledge how many (if any) language technologists are needed in IT companies.
* Results (especially language resources) are often subject to copyright protection, that is the reason why it is difficult to make results available and to license results. Question is how to minimize the burden of these managements.


Thank you!
www.keeletehnoloogia.ee
ee.clarin.eu
CELR
Baltic HLT 2016, Riga
06.10.2016