Strategic Importance of Language Technology in Estonia

Kadri Vider
Kadri.vider@ut.ee

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

CELR

Euroopa Liit
Euroopa
Regionaalarengu Fond

Eesti tuleviku heaks

Estonian language

- Finno-Ugric language, close to Finnish, far from Hungarian
- Ca 1 million mother-tongue speakers all over the world
- Only official language in Republic of Estonia

“A language is a dialect with an army and navy”

? … but what about dialects with LT support?

CELR

Estonian as official language

- Official language in European Union
- Language Act of Republic of Estonia
  - § 3. Status of Estonian language
  - § 4. Official and public use of language and Estonian Literary Standard
  - § 5. Foreign language and language of national minorities
  - § 7. The Estonian Language Council
- Development Plan of the Estonian Language (2011-2017)
  - chpt 4. Study of Estonian and language resources
  - chpt 5. Language-technological support of the Estonian language

CELR

Larger ethnic groups in Estonia

Population and Housing Census in 2011

Total population: 1,3 million

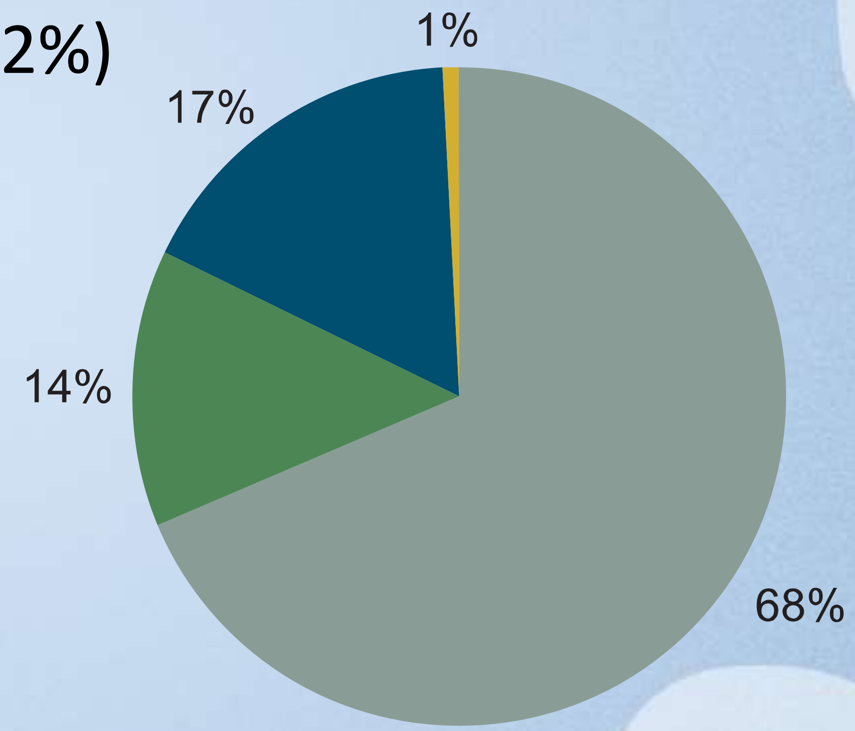
69,7% Estonians

25,2% Russians

2011	
Kogurahvastik *Total population*	1 294 455
Eestlased *Estonians*	902 547
Venelased *Russians*	326 235
Ukrainlased *Ukrainians*	22 573
Valgevenelased *Byelorussians*	12 579
Soomlased *Finns*	7 589
Tatarlased *Tatars*	1 993
Juudid *Jews*	1 973
Lätlased *Latvians*	1 764
Leedulased *Lithuanians*	1 727

Speakers of Estonian language

- Estonian as mother tongue - 68%
  - – 15.4% of them are also able to speak a dialect: **Võru or Setu** dialect (11,2%) insular dialect (2,8%)
- Estonian as foreign language – 14%
- Do not speak Estonian – 17%



NPELT = National Programme for Estonian Language Technology

State-funded activities in LT since 2006

Main objective from the Development Plan of the Estonian Language (2011-2017):

"…language technology support for the Estonian language will be on an equal level with that of other languages in countries with advanced language technology."

EESTI KEELETEHNOLOOGIA
NATIONAL PROGRAMME FOR ESTONIAN
LANGUAGE TECHNOLOGY

- www.keeletehnoloogia.ee
- Financed from state budget
  - 2006-2010: total 3,4 M€
  - 2011-2017: **765 K€** per year
- Managed by Steering Committee
- Results of projects are declared public. Center of Estonian Language Resources (CELR) is required to deposit all such resources and tools for preservation and long term access.

CELR

NPELT 2006-2010

- HLT-related R&D activities including the creation of reusable language resources and development of essential linguistic software (up to the working prototypes) as well as bringing the relevant language technology infrastructure up to date.
- 3 measures:
  - Software prototypes for LT
  - Language resources
  - Center of Estonian Language Resources
- 33 projects in wide range of topics:
  - speech synthesis and recognition,
  - compilation of digital language corpora,
  - lexical resources and tools such as database of an Estonian–X dictionary and lexicographer’s workbench,
  - machine translation,
  - information dialogue,
  - morphosyntactic and semantic analysis,
  - language software for the Web

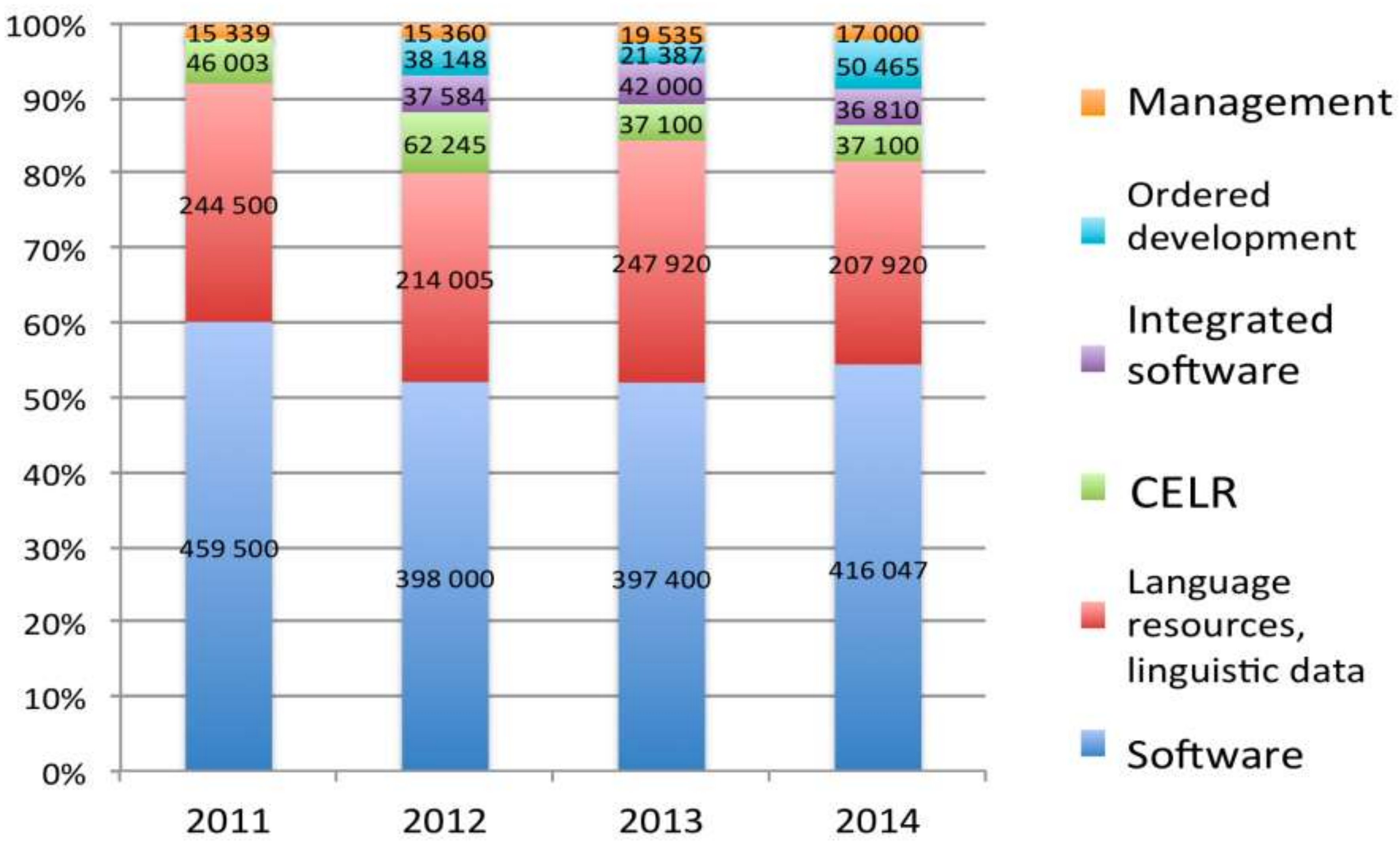
NPELT 2011-2017

- focus on the implementation and integration of the existing resources and software prototypes in public services.
- 3 old and 2 new measures:
  - 1. R&D to create new LT software;
  - 2. Projects for new language resources;
  - 3. Center of Estonian Language Resources (CELR);
  - 4. Projects for integrating language-specific software into other applications;
  - 5. Specially aimed projects (on the order of Steering Committee)

⇒ **all results (software and linguistic data) must be (re-)usable as freely as possible!**

CELR

Financing of NPELT in 2011-2014 (**765 K€** per year)



1. R&D projects building LT software (prototypes)

- Range of results is quite broad: from software prototypes to component software, not as often applications for end users.
- Some impressive projects:
  - Speech recognition project building web and mobile applications (TTU Institute of Cybernetics).
  - “Application Suite for voicing and broadcasting subtitles on television” (Institute of Estonian Language).

2. R&D projects building language resources (and LT data)

- provides digital language data (incl corpora, lexicons) for everyday language user and for language research as well as to IT-applications.
- projects also relate to topics of **big data** and **digital cultural heritage**.
- Some remarkable results:
  - Keeleveeb portal (www.keeleveeb.ee) - linked corpora, lexicons, termbases and NLP tools for end-users
  - Estonian WordNet

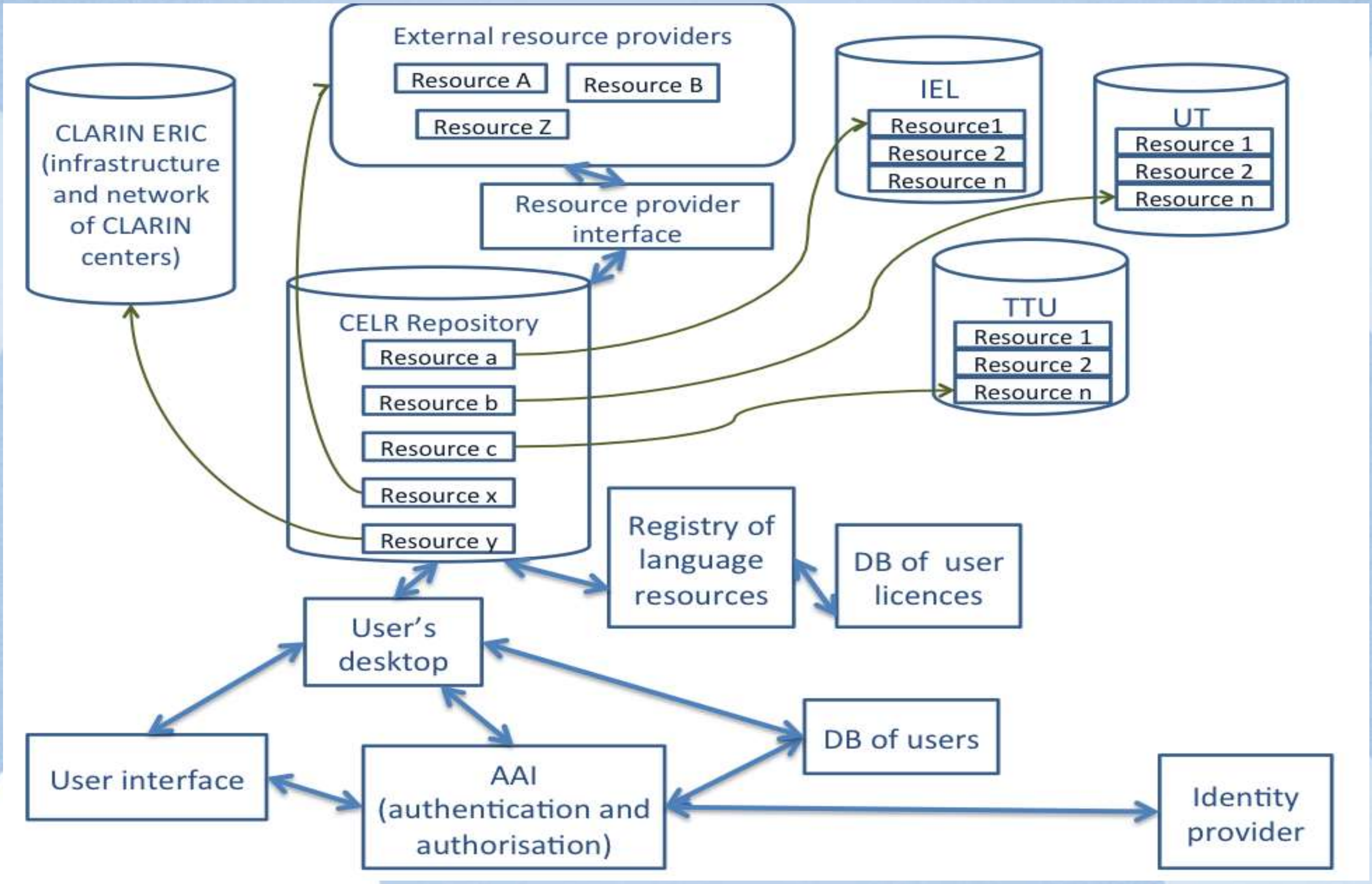
Center of Estonian Language Resources (CELR)

- The research infrastructure of Estonian language digital resources
  - language data: dictionaries, corpora (both text and speech),
  - language technology tools (software)

- To **make available** to everyone working with digital language materials
- To coordinate and organise the **documentation** and **archiving** of the resources
- To develop and promote language technology **standards**
- 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.**

Performance and structure of CELR



4. Integrated language software
5. Development projects ordered by SC

4. Technical aids for people with special needs and interfaces for public services are expected.

- notable project „Generation of Audiobooks and voicing interface of Digar“ (Institute of Estonian Language)

5. Software development projects ordered on the proposal of the steering committee of NPELT

- open-source morphological analyzer software was ordered from the only one Estonian fully language technology company Filosoft (www.filosoft.ee).

CELR

Risks in NPELT to achieve the objectives

- Projects are applied within an open competition – ideas which inspire researchers do not fully cover the objectives -> the language technology support for Estonian is not systematically developed;
- R&D 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 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
- Results (especially language resources) are often subject to copyright protection -> problems to make results available

CELR

Critical factors and risks of NPELT

- Underfunding may cause especially underdevelopment of software and language technology solutions
- Number of specialists working in the field is limited and salaries in business sector are higher than in academia
- Relatively low interest of Estonian IT-enterprises to develop LT software
- Potential change in language attitudes and use of software English versions
- Fetishising LT leads higher expectations to LT solutions than they can offer

LT support in Estonia

- Continuous funding and education have paid off
  - State of LT relatively good compared to languages with same amount of speakers.
- Estonian language is not commercially viable.
  - Only a few small enterprises active in LT
  - State needs to fund and coordinate the creation of basic resources and technologies
- Public funding
  - Results publicly available
  - LT Roadmaps and strategies – concentrate on the resources that we need.
- Research needs not only funding, but people as well
  - 2 bachelor programs, 2 master programs, 2 doctoral schools, 2 universities
  - LT researchers with base from computer sciences or linguistics

LT support for Estonian

- State of language resources is **reasonably good**
  - Public funding has guaranteed availability
  - Resources have been built for decades
  - More advanced resources lacking – syntactic, semantic annotatation, multimedia corpora
- More advanced tools are lacking, as is text generation.
- **Commercial tools** for text analysis are of high quality, but not freely available.
  Freely available (and publicly funded) tools are of reasonably good quality, but need further development.
- Research has resulted in tools that are available, but should be regarded as **prototypes** – they can and should be **further developed** into widely usable tools.
- As a result of research projects we have high-quality software, but **sustainability** can be an issue
  - Lack of documentation, standardisation

Thank you for attention!

Tänan tähelepanu eest!