EESTI KEELE KEELETEHNOLOOGILINE TUGI
NATIONAL PROGRAMME FOR ESTONIAN LANGUAGE TECHNOLOGY

National Programme for Estonian Language Technology: a Pre-final Summary

Einar Meister, Jaak Vilo* & Neeme Kahusk*****

Vice-chairman, *Chairman & * Coordinator of the Programme

Outline

HLT evolution in Estonia

Management

Financing

Supported projects

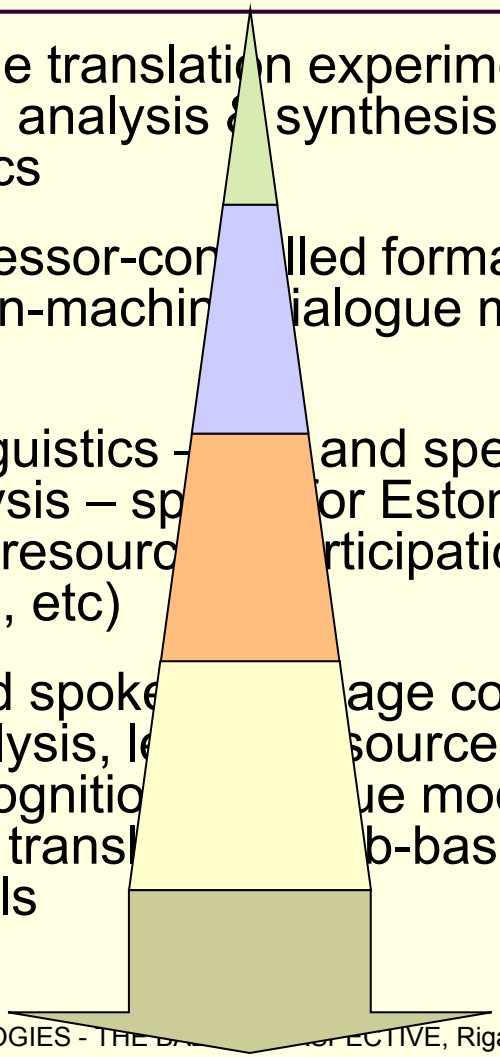
Research groups

Future prospects

Summary

HLT evolution in Estonia

- **1960-70s:** machine translation experiments, experimental phonetics, speech analysis & synthesis, semantic analysis, computer linguistics
- **1980s:** microprocessor-controlled formant synthesis, speech recognition, human-machine dialogue modelling, electronic dictionaries
- **1990s:** corpus linguistics – text and speech corpora, morphologic analysis – spell-checker for Estonian, electronic dictionaries, Web-resources, participation in EU-projects (WordNet, BABEL, etc)
- **2000s:** written and spoken language corpora, morpho-syntactic and semantic analysis, lexical resources and tools, speech synthesis and recognition, dialogue models, information retrieval, machine translation, web-based access to different resources and tools



HLT evolution in Estonia

Coordinated actions:

- **Estonian HLT program** supported by the Estonian Informatics Centre (1997-2000)
- **EU FP5 project eVikings II** (2002-2005): Roadmap for Estonian HLT 2004-2011
- **Centre of Excellence in HLT** (2003): successful in first round, failed in final round
- **Estonian Language Technology Development Centre** (2005): accepted for financing, but failed due to the withdrawal of the main industrial partner
- **National programme "Estonian Language and Cultural Heritage"** (1999-2003): some HLT-projects funded
- **National programme "Estonian Language and National Memory"** (2004-2008): sub-programme for Estonian HLT (2004-2005)
- **Development Strategy of the Estonian Language 2004-2010**
- **National Programme for Estonian Language Technology (2006-2010)**

National Programme for Estonian Language Technology 2006-2010

Government supported funding initiative aimed at developing of Estonian language resources and language-specific software in order to enable Estonian to function in the modern information technology environment

EESTI KEELE KEELETEHNOLOOGILINE TUGI
NATIONAL PROGRAMME FOR ESTONIAN LANGUAGE TECHNOLOGY

Estonian Ministry of Education and Research

Management (1)

- **Steering committee** of 9 members including representatives of the ministries and HLT-experts responsible for:
  - evaluation of project proposals and progress reports
  - making funding proposals
  - purposeful use of public funding
  - surveying the developments in the HLT field on the national and international scale

Management (2)

- **Programme coordinator** responsible for:
  - preparing calls for projects
  - project contracts and reports
  - communication between the ministry, steering committee and project leaders
  - documentation and Web-site administration

Management (3)

- **General rules:**
  - financing of projects based on open competition
  - evaluation of projects based on well-established criteria
  - international standards/formats need to be followed
  - groups are requested to provide annual progress reports
  - developed prototypes and language resources are public

Management (4)

- **Project evaluation criteria:**
  - for new applications:
    - relevance of the proposal in the context of the programme
    - methods applied to achieve the goals of the project
    - competence and experience of the project team
    - usefulness of project's results for other projects
    - compatibility and use of standards
    - etc.
  - for assessment of the annual progress of on-going projects

Funding (1)

- **Funding decision is based on the average score of individual ratings given by the steering committee members**

Average score	Coefficient
90-100%	0,8-1
65-90%	0,7-0,9
< 65%	0

Depending on available funding and number of applications

- Ca 33% for corpus projects, 65% for software & research projects, 1-2% for management

Statistics: projects & funding

	2006	2007	2008	2009	2010
Number of project applications	**22**	**22** (18+4)	**23** (20+3)	**24** (15+9)	**24** (22+2)
Number of funded projects	**18**	**20** (18+2)	**23** (20+3)	**23** (15+8)	**24** (22+2)
Total funding, MEEK (MEUR)	**7.3** (0.47)	**7.1** (0.46)	**13.4** (0.86)	**12.9** (0.83)	**11.8** (0.75)

Projects
http://www.keeletehnoloogia.ee/projects

- **Speech corpora** – emotional speech, spontaneous speech, dialogues, L2 speech, radio news and talk shows

- **Text corpora** – written language corpus, multi-lingual parallel corpora, resources for interactive language learning

- **Research/technology development** – speech recognition & synthesis, machine translation, information retrieval, lexicographic tools, syntactic & semantic analysis, dialogue modeling, rule-based language software, intelligent search engine, variations in speech production and perception

Key players (1)

- **University of Tartu:**
  - morphology, syntax, semantics, and machine translation
  - corpora of written and spoken language, dialogue corpora, parallel corpora, lexical and semantic database (thesaurus, Estonian WordNet), phonetic corpus of spontaneous speech
  - rule-based language software, information retrieval, interactive Web-based language learning

Key players (2)

- **Institute of the Estonian Language:**
  - Corpus-based speech synthesis for Estonian
  - Estonian Emotional Speech Corpus
  - Lexicographer's workbench

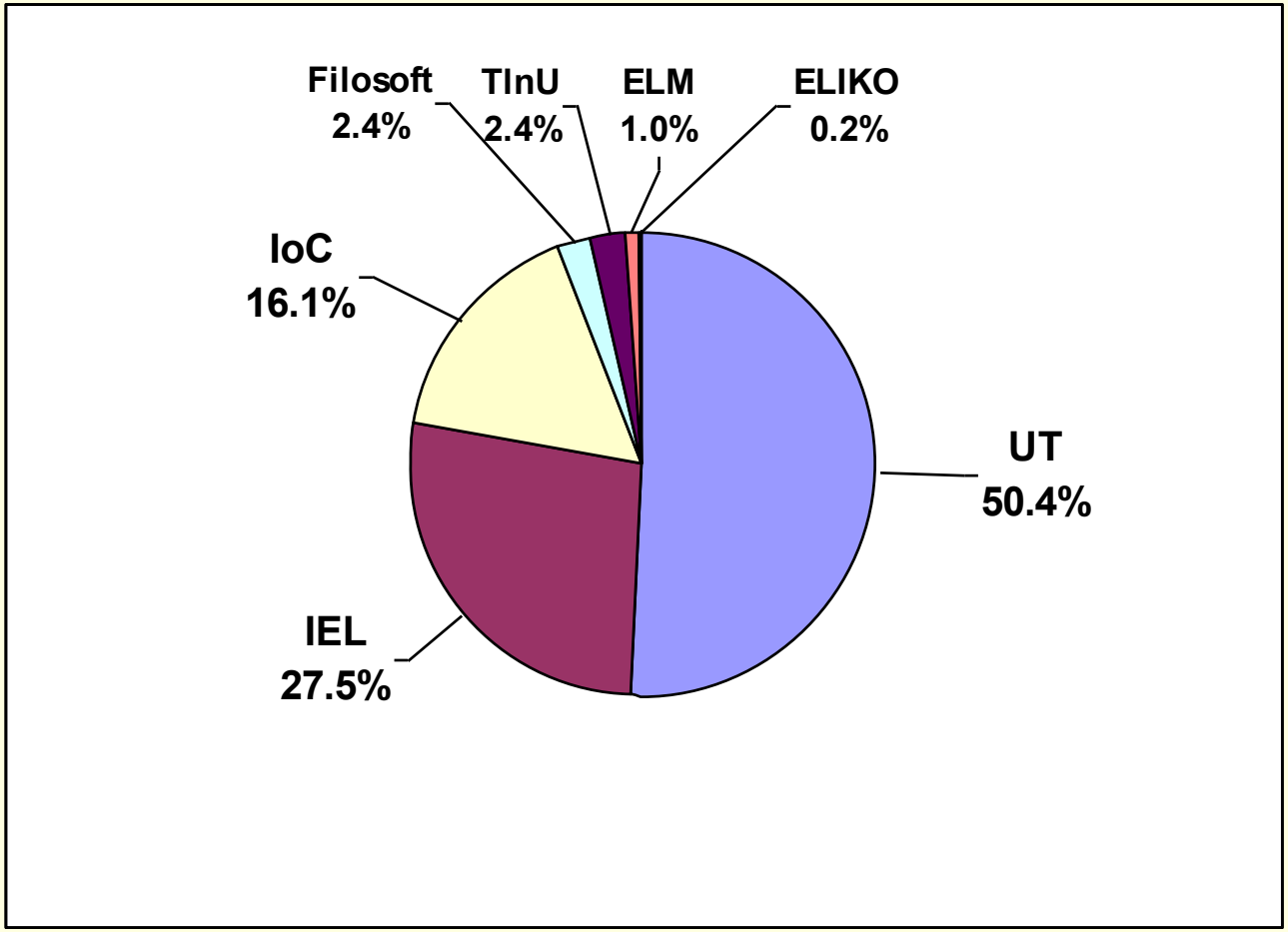
Key players (3)

- **Institute of Cybernetics at Tallinn University of Technology:**
  - automatic speech recognition in Estonian
  - variability in speech production and perception
  - speech corpora including radio news and talk shows, lecture speech, foreign-accented speech

Key players (4)

- **Filosoft:** corpus query in the Estonian language website keeleveeb.ee
- **Tallinn University:** Estonian Interlanguage Corpus
- **Estonian Literary Museum:** electronic dictionary of idiomatic expressions
- **ELIKO:** a prototype of Controlled Natural Language module for knowledge-based systems

Division of funding 2006-2010



Distribution of results (1)

- **Centre of Estonian Language Resources:**
  - the project launched in 2008 at the University of Tartu
  - partners – Institute of the Estonian Language and Institute of Cybernetics at TUT
  - main goal – to develop the infrastructure for archiving, documenting and distribution of Estonian language resources and software tools
  - cooperation with CLARIN project
  - in 2010 included into the Estonian Research Infrastructures Roadmap

Distribution of results (2)

- **Programme conferences:**
  - 1st conference: November 2007, Tallinn
  - 2nd conference: April 2009, Tartu
  - **3rd conference: November 25-26, 2010, Tartu**

Supporting activities

- **Development of human resources:**
  - Doctoral School of Linguistics and Language Technology (2005-2008)
  - Doctoral School in Information and Communication Technologies (2009-2015)
  - Centre of Excellence in Computer Science (2008-2015)
  - Curricula on computer linguistics and language technology at the University of Tartu
  - Speech technology course at Tallinn University of Technology

Future prospects

- **Currently under development:**
  - Estonian BLARK
  - Estonian HLT Roadmap for 2011-2017
  - follow-up programme for 2011-2017
- Focus of the follow-up programme on resources, software tools and **integrated prototypes for public applications**
- **Important issues:**
  - availability of resources and tools via Centre of Estonian Language Resources
  - promoting HLT integration into public and commercial applications
  - urgent need for HLT-engineers and researchers

Summary

- The national programme has created favourable conditions for HLT development in Estonia
- < 50 MEEK (3.5 MEUR) invested into HLT area, < 30 different projects funded
- Remarkable progress in the amount and diversity of Estonian language resources and tools
- Good bases for future applications and international cooperation
- Estonian HLT will be not ready by the end of 2010 – a follow-up programme is necessary

Last, but not least…

Steven Krauwer's talk at the 2nd Baltic HLT conference in Tallinn 2005:

"How to survive in a multilingual EU?“

- *Do not expect too much from the EU due to the* subsidiarity principle
- *National level activities are important – if you don’t care of your language no one will do!*
- *There are at least two areas which should be evolved mainly at the national level – creation of language resources and training of languages technologists*

Really final…

Are we moving fast enough?

Interspeech 2010:

- *Real time speech-to-speech translation*
- *Google voice browser*, etc