Strengthening the Estonian Language Technology

Einar Meister[1], Jaak Vilo[2]
[1]Institute of Cybernetics at Tallinn University of Technology
Akadeemia tee 21, 12618 Tallinn, Estonia
[2]Department of Computer Science
University of Tartu
J.Liivi 2, 50409 Tartu, Estonia
E-mail: einar@ioc.ee, jaak.vilo@ut.ee

Abstract

The paper will give an overview of developments in Estonia in the field of Human Language Technologies. Despite of the fact that Estonian is one of the smallest official languages in EU and therefore in less favourable position in the HLT-market, the national initiatives are undertaken in order to promote HLT development in Estonia. The paper will introduce recent activities in Estonia, including National Programme for Estonian Language Technology (2006-2010).

1. Introduction

The development efforts of human-computer interaction during the past few decades have been directed towards natural communication using spoken language input and output. For several, especially "big" languages, progress in language technology has been impressive – research results have been successfully exploited in commercial products and services, and the HLT-market shows growing trends. According to the Euromap report (Joscelyne, Lockwood, 2003) on HLT progress in EU countries, the leading positions are held by the UK, Germany, France, the Netherlands and Finland. In the case of the first three countries it can be explained mainly by large market demands, whereas in the latter cases the leading position has been achieved due to several simultaneous factors – healthy environment for R&D, relatively large and strong research community and significant national-level support in the HLT area.

Although linguistic and cultural diversity are the core values of the EU and discrimination based on language is prohibited by the EU's charter of fundamental rights (article 22) we need to face the fact that there are primary, secondary and even tertiary languages of commercial relevance (TC-STAR report, 2006). Development of HLT tools for a new language is a more or less fixed effort and does not correlate with the number of speakers; therefore the smaller languages are in less favourite position, as the costs per capita for HLT development will be higher.

What should be done for smaller languages in order to strengthen their market positions and survival in a multilingual EU? – these are crucial questions for smaller countries and also for EU language policy makers wanting to prevent Gutenberg's effect from taking place in the computer age. These issues have been addressed in Krauwer's papers (2005, 2006). Krauwer's claim that the strong industrial bias of EU programmes has led to the situation where the major part of HLT funding is used to support a few major EU languages seams to hold true. As there are not many options (due to the subsidiarity principle) to get financial support from the EU for the technological development of smaller languages, activities on the national level are of great importance.

In Estonia several activities to promote R&D in HLT area have been undertaken during the last decade. Mostly these activities have been initiated by the academic groups working on HLT-related topics; in parallel with academic research a lot of effort has been put into explaining the role of HLT in the information society. Although not all initiatives were fully successful, they played an enlightening role among decision-makers and contributed to the forming of a positive attitude in the society. As a result of the joint effort of researchers and the Ministry of Science and Education, the National Programme for Estonian Language Technology (2006-10) was launched.

In this paper we will share our experiences in promoting HLT-related national activities and introduce the Estonian HLT roadmap as well as on-going R&D projects.

2. HLT research in Estonia

The history of HLT research in Estonia dates back to the 1960s when the first academic groups working on computer linguistics, experimental phonetics and speech analysis were established in Estonia. After 1991, when Estonia re-established its independence, the whole system of research structure in the country was reorganised and new financing schemes were introduced. Most of today's HLT research units have sprung up from these former groups.

There are three key players working in the field of HLT in Estonia:

(1) **University of Tartu,** represented mainly by the **Research Group on Computer Linguistics** (http://www.cl.ut.ee). Their research areas cover:
– formal descriptions of morphology, syntax and semantics of the Estonian;
– creating Estonian language resources: electronic corpora of written and spoken language, dialogue corpora, parallel corpora, lexical and semantic database (thesaurus, Estonian WordNet);

– software development for morphological, syntactic and semantic analysis and synthesis.

In addition, two further groups (bioinformatics and phonetics) contribute to HLT field.

(2) **Institute of the Estonian Language, Research Group on Language Technology** (http://www.eki.ee), focused on:

– rule-based morphological systems: formal grammars and software (morphological synthesis and analysis, morphological disambiguation);

– language resources: electronic versions of traditional dictionaries, linguistic databases, text-based dictionaries, lexicons for machine translation, www-applications;

– phonetics and speech technology: text-to-speech synthesis (TTS) and linguistic problems (modelling of speech prosody, relations between syntax and prosody) and speech databases.

(3) **Institute of Cybernetics at Tallinn University of Technology** represented by the **Laboratory of Phonetics and Speech Technology** (http://wwww.phon.ioc.ee). It's R&D activities include:

– experimental phonetics: research on Estonian sound system and prosody including Estonian as L2;

– speech technology: speech analysis and speech synthesis, automatic speech recognition (ASR);

– speech databases: Estonian BABEL, Estonian SpeechDat, etc.

There also exist a few small private HLT companies:

Filosoft (http://www.filosoft.ee) – a spin-off company of Tartu University established in 1993, provider of several software products (speller, hyphenator and thesaurus for Estonian, speller and hyphenator for Latvian) and dictionaries for several platforms (MS Windows, Mac OS X, Unix). The company runs the language portal Keeleveeb (http://www.keeleveeb.ee) offering free access to different on-line dictionaries, software and corpora.

Keelevara (http://www.keelevara.ee) was founded in 2004 in order to provide on-line access to several professional electronic dictionaries and lexicons, access to some dictionaries is free.

Tilde Eesti (http://www.tilde.ee) is a branch of Latvian company Tilde (http://www.tilde.lv), established in 1991. Tilde's products cover localized fonts, Latvian and Lithuanian language support, proofing tools, electronic dictionaries, multimedia products, etc. Tilde Eesti is focused on software localisation and translation services.

TEA Publishers (http://www.tea.ee) **–** established in 1991, one of the leading publishers of economics dictionaries and foreign language textbooks in Estonia.

Imprimaatur – founded in 1996, offers consulting, training and quality assurance services related to translation and term banks.

Festart – established in 1995, provider of electronic dictionaries English <–> Estonian, Russian <–> Estonian.

Nekstom – OCR for Estonian, distributor of ABBYY software in Estonia.

2.1 HLT financing

Reforms of research funding in the beginning of the 1990s mark a new era for the academic community in Estonia. A competition-based funding scheme was introduced where all research fields had to compete for survival. HLT research groups survived quite well due to successful participation in several international projects (e.g. EU Copernicus). Starting at the end of the 1990s, additional funding sources were opened:

– the Estonian Language Technology programme initiated by the Estonian Informatics Centre (1998-2000). Within this programme the first Development Plan for Estonian Language Technology was compiled in 1999;

– the national programmes “Estonian Language and Cultural Heritage" (1999-2003) and “Estonian Language and National Memory” (2004-2008) including sub-programmes for HLT.

HLT key-players were involved also in EU FP5 project “eVikings II: Establishment of the Virtual Centre of Excellence for IST RTD in Estonia” (2002-2005). One important outcome of the project was the Estonian HLT Roadmap for 2004-2011. Within this project also two further applications (for the Estonian Language Technology Competence Centre and for the Centre of Excellence in HLT) were submitted to different funding bodies in 2003. Both applications were not fully successful, but they played an important role in paving the way to the national HLT programme.

3. Estonian HLT Roadmap

The roadmap (Figure 1) compiled in 2004 shows the baseline – the resources and tools developed in Estonia during several years before 2004, and presents the future developments in three major action lines:

Action Line 1: Spoken Language Technology including:

– speech synthesis: creating Estonian TTS software and development of an audio-visual synthesis prototype;

– speech recognition: creating a prototype of limited vocabulary ASR and development of language-specific methods for unlimited vocabulary ASR;

– dialogue systems: creating limited-domain intelligent services capable of replacing routine human work.

Action Line 2: Written Language Technology including:

– language processing methods: formalisms for automated processing of different language levels (morphology, syntax, semantics, pragmatics), modelling and creating of corresponding prototypes;

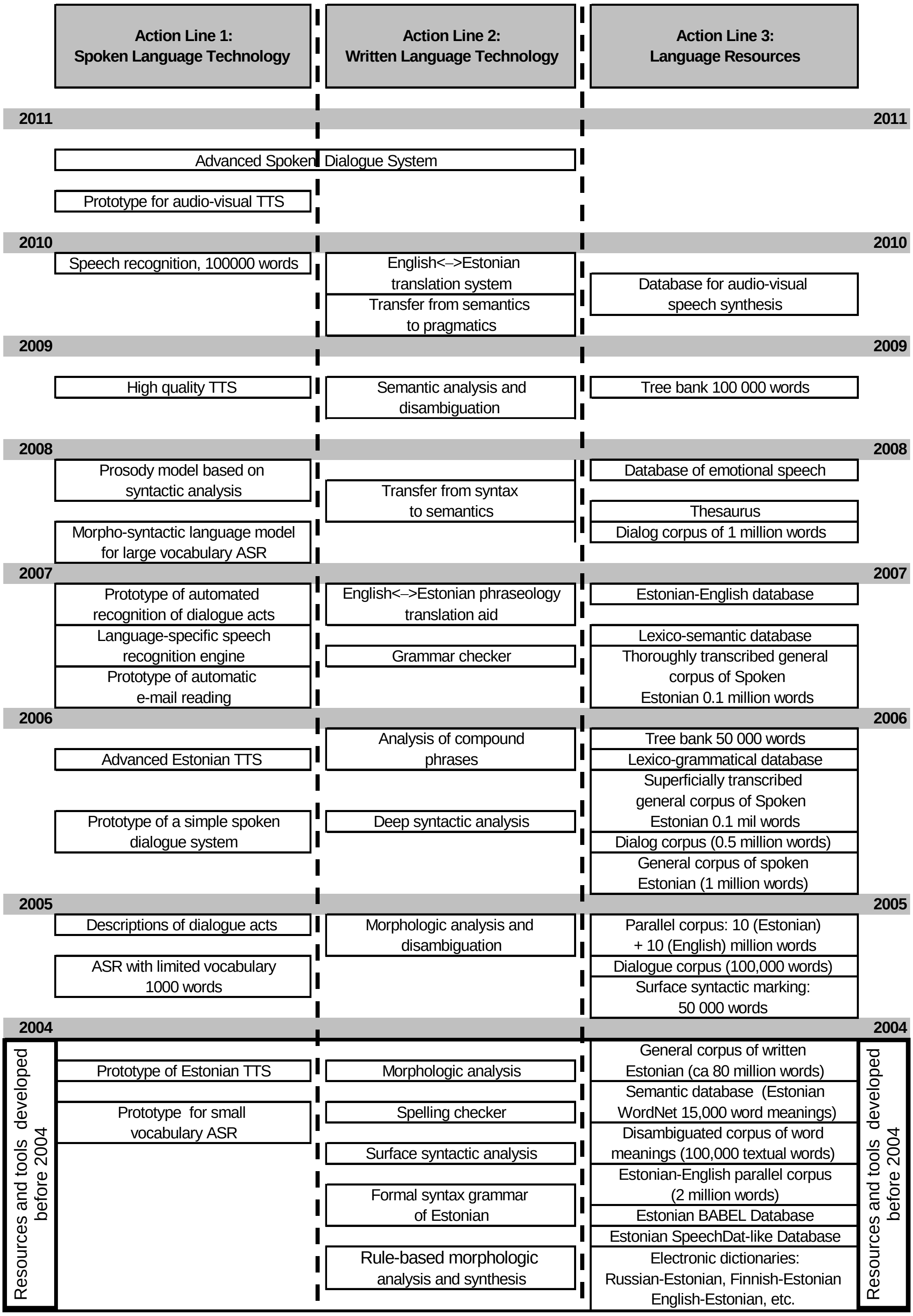
– machine translation: create methods for translating to and from Estonian, compile multilingual vocabularies and mechanisms of transforming syntactic structures; develop a prototype for Estonian <–> English machine translation.

Action Line 3: Language Resources including:

– creating infrastructure for collection and management of different language resources;

– collecting different types of resources: speech and text corpora, and electronic dictionaries.

Comparing the roadmap to the achievements in 2008 we can see good progress in all action lines, nevertheless an update of the roadmap is necessary.



	Action Line 1: Spoken Language Technology	Action Line 2: Written Language Technology	Action Line 3: Language Resources	
2011				2011
	Advanced Spoken Dialogue System	Advanced Spoken Dialogue System		
	Prototype for audio-visual TTS			
2010				2010
	Speech recognition, 100000 words	English<–>Estonian translation system Transfer from semantics to pragmatics	Database for audio-visual speech synthesis	
2009				2009
	High quality TTS	Semantic analysis and disambiguation	Tree bank 100 000 words	
2008				2008
	Prosody model based on syntactic analysis Morpho-syntactic language model for large vocabulary ASR	Transfer from syntax to semantics	Database of emotional speech Thesaurus Dialog corpus of 1 million words	
2007				2007
	Prototype of automated recognition of dialogue acts Language-specific speech recognition engine Prototype of automatic e-mail reading	English<–>Estonian phraseology translation aid Grammar checker	Estonian-English database Lexico-semantic database Thoroughly transcribed general corpus of Spoken Estonian 0.1 million words	
2006				2006
	Advanced Estonian TTS Prototype of a simple spoken dialogue system	Analysis of compound phrases Deep syntactic analysis	Tree bank 50 000 words Lexico-grammatical database Superficially transcribed general corpus of Spoken Estonian 0.1 mil words Dialog corpus (0.5 million words) General corpus of spoken Estonian (1 million words)	
2005				2005
	Descriptions of dialogue acts ASR with limited vocabulary 1000 words	Morphologic analysis and disambiguation	Parallel corpus: 10 (Estonian) + 10 (English) million words Dialogue corpus (100,000 words) Surface syntactic marking: 50 000 words	
2004				2004
Resources and tools developed before 2004	Prototype of Estonian TTS Prototype for small vocabulary ASR	Morphologic analysis Spelling checker Surface syntactic analysis Formal syntax grammar of Estonian Rule-based morphologic analysis and synthesis	General corpus of written Estonian (ca 80 million words) Semantic database (Estonian WordNet 15,000 word meanings) Disambiguated corpus of word meanings (100,000 textual words) Estonian-English parallel corpus (2 million words) Estonian BABEL Database Estonian SpeechDat-like Database Electronic dictionaries: Russian-Estonian, Finnish-Estonian English-Estonian, etc.	Resources and tools developed before 2004

Figure 1. Estonian HLT Roadmap for 2004-2011

4. Towards national HLT programme

In 2003 the Development Strategy of the Estonian Language 2004-2010 was compiled by the members of the Estonian Language Council and was approved by the Estonian Government on August 5, 2004.
http://www.eki.ee/keelenoukogu/strat_en.pdf
The strategy provides a research-based description of the situation of the Estonian language, the objectives that need to be achieved, the necessary steps and institutions and people involved. The development plan of the Estonian language covers all the major areas of language use including language technology.

4.1 National Programme for Estonian Language Technology (NPELT)

NPELT (http://www.keeletehnoloogia.ee) was compiled in 2005 by a group of HLT experts and launched by the Ministry of Science and Education in 2006 for a period of five years (2006-2010).

The main goal of NPELT is to develop technology support for the Estonian language to the level that would allow functioning of Estonian in the modern information society. NPELT is funding HLT-related R&D activities including 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. The resources and prototypes funded by the national programme are declared public.

NPELT management is carried out by a steering committee of 9 members (including HLT experts and representatives of the ministries), and a programme coordinator. Responsibilities of the steering committee include the 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, etc. General rules adopted by the committee:

- financing of projects based on open competition,
- groups are requested to provide annual progress reports,
- evaluation of projects based on well-established criteria,
- international standards/formats need to be followed,
- access to the developed prototypes and language resources should be free or based on licence agreements.

Financing of the programme: ca 0.5 M€ per year in 2006 and 2007, ca 1.1 M€ per year for 2008 – 2010, of which about 33% should be used for the creation of language resources, 66% for research and software development, and 1% for the programme management.

On-going projects: In 2008, 23 projects have been funded (2006: 17, 2007: 20) which cover a wide range of topics (see http://www.keeletehnoloogia.ee/projektid):

- speech corpora: emotional speech, spontaneous speech, dialogues, L2 speech, etc;
- text corpora: written language corpus, multi-lingual parallel corpora, etc.
- research/technology development – speech recognition, speech synthesis, machine translation, information retrieval, lexicographic tools, syntactic analysis, semantic analysis, dialogue modelling, variations in speech production and perception, etc.

5. Doctoral School in Linguistics and Language Technology

To improve the quality of doctoral studies in linguistics and language technology and to meet the growing need for HLT experts the Doctoral School of Linguistics and Language Technology (DSLLT, http://www.fl.ut.ee/kttdk) was launched at Tartu University for 2005 to 2008. Partners of the DSLLT are the above-mentioned HLT key players and several foreign universities as well as some local private companies. The activities of the school have strongly contributed to the effectiveness of doctoral studies of many students; about 10 PhD theses in HLT areas have been prepared and defended with DSLLT support.

6. Conclusions and future prospects

The national programme has created favourable conditions for HLT development in Estonia. Obviously not all HLT fields are equally addressed and it would be naive to expect that all essential prototypes and resources will be created within a short period.

The steering committee is planning an update of the HLT roadmap and takes the initiative towards defining a BLARK (Basic Language Resource Kit) for Estonian. A new project call for 2008 is under preparation in order to attract IT companies to implement the existing prototypes. Our national activities synchronize well with our participation in the EU CLARIN project.

7. Acknowledgements

We are grateful to the members of the steering committee, HLT-experts involved in the preparation of NPELT and the Ministry of Science and Education for their enormous effort in making the national HLT-programme real. All researchers of on-going projects are acknowledged for their devoted work.

8. References

Joscelyne, A., Lockwood, R. (2003). Benchmarking HLT progress in Europe. The EUROMAP Study. Copenhagen 2003.

Krauwer, S. (2005). How to survive in a multilingual EU? *Proc. of The Second Baltic Conference on HLT*, April 4-5, 2005, Tallinn, Estonia, pp. 61-66.

Krauwer, S. (2006). Strengthening the smaller languages in Europe. *Proc. of 5th Slovenian and 1st International Language Technologies Conference,* October 9-10, 2006, Ljubljana, Slovenia. Retrieved on 11/6/2007 from http://nl.ijs.si/is-ltc06/proc/01_Krauwer.pdf

TC-STAR report (2006). Human language Technologies for Europe. Retrieved on 10/12/2007 from http://www.tc-star.org/pubblicazioni/D17_HLT_ENG.pdf