Einar Meister

Human Language Technology developments in Estonia

Lühikokkuvõte

Eesti keel on üks väiksema kõnelejate arvuga keeli Euroopa Liidus ja seetõttu on keeletehnoloogiline arendustöö eesti keele jaoks majanduslikult ebaotstarbekas, kuid keele tuleviku seisukohast äärmiselt oluline. Keeletehnoloogia arendamiseks Eestis on käivitatud mitmeid ettevõtmisi. Uurimistöid alustati juba 1960ndatel aastatel ja tänaseks on keeletehnoloogia saavutanud tunnustatud positsiooni. Artikkel annab ülevaate keeletehnoloogia arengust Eestis keskendudes peamiselt riiklikule programmile "Eesti keele keeletehnoloogiline tugi (2006-2010)". Tutvustatakse programmi eesmärke, ülesehitust ja juhtimist ning mitmete projektide tulemusi. Lisaks käsitletakse aastateks 2011-2017 kavandatud keeletehnoloogia jätkuprogrammi, spetsialistide järelkasvu ning teadustulemuste rakendamisega seotud probleeme.

Abstract

Estonian is one of the smallest official languages in the EU and therefore in a less favourable position on the Human Language Technologies (HLT) market, although this is extremely important for survival of the language. To promote HLT developments in Estonia national initiatives have been undertaken. HLT research in Estonia was started in the early 1960s and by today has gained a recognized position. The paper gives an overview of developments in the field of HLT focussing on the National Programme for Estonian Language Technology (2006-2010). The management of the programme and projects covering different areas of language technology are introduced. In addition, the follow-up programme for 2011-2017 and the issues of human resources and cooperation with industry are discussed.

1. Introduction

Linguistic and cultural diversity are core values of the European Union protected by EU legislations as well as promoted by several funding instruments. The report "Human Language Technology for Europe" compiled within the TC-STAR project states that for Europe, Human Language Technology (HLT) is an economic, political and cultural necessity since the European Union is a multilingual society by design (Lazzari 2006). Despite the fact that all EU languages are declared equal, however, as the report states, there are primary, secondary and even tertiary languages of commercial relevance; especially languages with a small number of speakers are at a disadvantage. Indeed, the official languages of the EU differ to a great extent from the point of view of existing technological support as well as availability and diversity of reusable language resources. This unbalanced situation is a result of multiple factors including the strength and number of academic HLT research groups in different countries, differences in national-level funding (both the public sector and industry) for research and technology development, as well as the disadvantageous funding practice of recent EU Framework programmes where most funding went to commercially attractive languages; in addition, the subsidiarity principle does not allow EU funding schemes to offer more favourable opportunities for HLT development of smaller languages (Krauwer 2005, 2006).

In recent years, several EU-level activities have been initiated in order to promote the development and wider use of HLT, for example, the CLARIN project (www.clarin.eu) for creating a Europe-wide infrastructure for common language resources; the FLaReNet

project (www.flarenet.ee) aiming at developing a common vision of the field of language resources and technologies and fostering a European strategy for consolidating the HLT sector; META-NET (www.meta-net.eu), a Network of Excellence building the Multilingual Europe Technology Alliance in order to join efforts towards furthering language technologies as a basis for the technological foundations of a multilingual European information society. Protecting the linguistic diversity and building the technological support for all official EU languages is certainly expensive (23 official languages, 506 language pairs) – the necessary investments should be shared with the European Commission and the Member States, in full agreement with the concept of "subsidiarity" (Mariani 2009).

In several EU countries diverse national level initiatives are undertaken in order to facilitate and coordinate research and development of HLT for national languages, e.g., in France (Mariani 2009), the Netherlands (Spyns/D'Halleweyn 2010; Odijk 2010), Sweden (Elenius et al. 2008), etc.; some effort has been made also for the languages without national state, e.g., Catalan (Melero et al. 2010).

In Estonia, the National Program for Estonian Language Technology (2006-2010) (NPELT) was launched in 2006. NPELT was a government supported funding initiative aimed at developing HLT for the Estonian language to the level that would allow functioning of Estonian in the modern information society. NPELT funded HLT-related R&D activities including the creation of reusable language resources and development of essential language-specific 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 program are declared public. In 2011, the follow-up program for Estonian Language Technology for the years 2011-2017 was approved.

The current paper gives an overview of the HLT developments in Estonia starting with a retrospect from the 1960s, then NPELT (2006-2010) will be introduced and finally the perspectives of the HLT developments within the follow-up program for 2011-2017 will be discussed.

2. HLT evolution in Estonia[1]

HLT development in Estonia can be characterized as an evolutionary process starting in the early 1960s when the first machine translation experiments were carried out and the analysis of legal texts using a computer was initiated at the University of Tartu. In the same decade two research units were established in Tallinn – the laboratory of experimental phonetics at the Institute of Estonian Language and the research group on speech analysis and synthesis at the Institute of Cybernetics.

In the 1970s studies on speech recognition and human-machine dialogue modelling were initiated and the transition to computer-based production of dictionaries was started. Experimental studies in Estonian phonetics and developments in speech analysis and synthesis techniques allowed the building of microprocessor-controlled formant synthesizers to begin.

[1] The provided brief overview of the HLT evolution in Estonian does not claim to be exhaustive and unbiased, it is rather the author's personal (insider) view of the main processes and development trends.

In the 1980s several text-to-speech systems for Russian and Estonian were developed at the Institute of Cybernetics and at the Institute of Estonian Language and exploited as output devices in automated control systems or as message readers for the blind. The research group on computer linguistics at the University of Tartu was formed with the main focus of study on morphology, syntax, semantics and human-machine dialogue.

After 1991, when Estonia re-established its independence, the whole system of academic research structure in the country was reorganised and new financing schemes were introduced. The restructuring of the political and economic system was accompanied by a remarkable decrease in the number of academic personnel – several researchers and engineers moved to governmental institutions and business, especially to the IT sector, where a large number of SMEs was established.

For the academic groups in the HLT area surviving the reforms, new opportunities for international cooperation opened up in the mid-1990s – Estonian research groups were able to join several EU projects such as EuroWordNet, BABEL, GLOSSER, TELRI, TELRI-II, etc. In addition to different corpus projects (both text and speech) carried out in the 1990s, a number of electronic dictionaries were made available via the Internet and a speller for Estonian was developed and commercialized by Filosoft Ltd (a spin-off company of Tartu University).

In the first decade of the 2000s HLT research in Estonian made substantial progress – the scope of research was remarkably broadened, involving areas such as morphologic, 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, and the amount as well as diversity of different reusable speech and text resources increased significantly.

The progress in Estonian HLT in the last decades was achieved through activities of academic groups who in parallel with academic research put a lot of effort into explaining the role of HLT in the information society and the need for developing language-specific resources and software. A number of concerted initiatives were successful and resulted in funding of different research projects from nationwide programmes or promoted international cooperation in the HLT field, for example:

– the Estonian HLT programme supported by the Estonian Informatics Centre (1997-2000); within this programme the first Development Plan for Estonian Language Technology was compiled in 1999;
– the EU FP5 project eVikings II (2002-2005) contributed to the development of the Roadmap for Estonian HLT 2004-2011 (Meister/Vilo 2008);
– the national programme "Estonian Language and Cultural Heritage" (1999-2003) funded some HLT projects;
– the national programme "Estonian Language and National Memory" (2004-2008) had a specific sub-programme for Estonian HLT (2004-2005).

Not all initiatives were fully successful, for example, the application for the Centre of Excellence in HLT (2003) was successful in the first round but failed in the final round, and the application for the Estonian Language Technology Development Centre (2005) was accepted for financing, but failed due to withdrawal of the main industrial partner.

However, all these initiatives played an enlightening role among decision-makers and contributed to the forming of a positive attitude in the society as well as paving the way for the national HLT programme.

In 2004 the Development Strategy of the Estonian Language 2004-2010 was compiled by the Estonian Language Council and approved by the Estonian Government. It involved a chapter on Estonian HLT which served as a base for the development of the National Programme for Estonian Language Technology. The programme for 2006-2010 was compiled by the joint effort of local HLT experts and the Ministry of Science and Education and approved in 2006.

3. National Programme for Estonian Language Technology (2006-2010)

The National Programme for Estonian Language Technology (NPELT) was a government-supported funding initiative aimed at developing Estonian language resources and language-specific software (up to working prototypes) in order to enable Estonian to function in the modern information technology environment. NPELT involved two main action lines:

Action line 1 for supporting projects of reusable language resource collection, including different **text corpora** (written language corpus, multi-lingual parallel corpora, resources for interactive language learning, etc) and **speech corpora** (emotional speech, spontaneous speech, dialogues, L2 speech, radio news and talk shows, etc).

Action line 2 for research of **methods** and development of **software prototypes** in a wide range of HLT areas such as speech recognition and synthesis, machine translation, information retrieval, lexicographic tools, syntactic and semantic analysis, dialogue modelling, rule-based language software, variations in speech production and perception, etc.

3.1 Steering committee

The management of the programme was carried out by a steering committee of 9 members including HLT experts and representatives of the ministries, and a programme coordinator. The steering committee was responsible for the evaluation of project proposals and progress reports according to established criteria, making funding proposals, surveying the developments in the HLT field on the national and international scale, etc. General rules adopted by the committee included the following:

- 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,
- the developed prototypes and language resources should be put in the public domain, only in exceptional circumstances access could be based on clear license agreements.

3.2 Project evaluation criteria

Two types of evaluation criteria were developed: (1) criteria for new project applications, and (2) criteria to assess the annual progress of on-going projects. The funding decision

of a new project was based on the average ratings of eleven features (sub-criteria) including the 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, whether the results of the project were useful for other projects, etc. In the case of ongoing projects the evaluation was based on annual progress reports which had to provide detailed information on how well the project had proceeded; objective measures were applied where possible (mainly in the case of resource projects).

3.3 Financing of the programme

The programme was financed out of the government budget, in 2006 and 2007 ca 0.5 M€ per year, ca 1.1 M€ in 2008, and ca 0.8 M€ per year in 2009 and 2010. According to the guidelines of the programme, ca 33% of total financing was used for projects focussed on the development of language resources, and ca 66% for research and software development; administration costs were limited to ca 1%.

3.4 Funded projects

The number of funded projects was slightly increased from year to year: 2006 – 17 projects, 2007 – 20 projects, 2008 and 2009 – 23 projects, and 2010 – 24 projects. Most of the projects were long-term projects spanning the years from 2006 to 2010, but also a few short-term projects (1-2 years) were funded. The projects covered a wide range of topics and were carried out mainly by three key players working in the field of HLT (see www.keeletehnoloogia.ee/projects):

1. University of Tartu, represented by three groups: (1) Research Group on Computer Linguistics, (2) Phonetics, and (3) Bioinformatics. Their projects were focused on:
   – 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.
2. Institute of the Estonian Language, represented by the Research Group on Language Technology, had three projects:
   – corpus-based speech synthesis for Estonian,
   – Estonian emotional speech corpus,
   – lexicographic tools.
3. Institute of Cybernetics at Tallinn University of Technology, represented by the Laboratory of Phonetics and Speech Technology, carried out three projects:
   – automatic speech recognition in Estonian,
   – variability issues in speech production and perception,
   – speech corpora including radio news and talk shows, lecture speech, foreign-accented speech.

In addition, there were other institutions and companies responsible for single projects:
– Tallinn University – Estonian Interlanguage Corpus,
– Estonian Literary Museum – electronic dictionary of idiomatic expressions,
– Filosoft – corpus query on the Estonian language website keeleveeb.ee,
– Eliko – prototype of a Controlled Natural Language module for knowledge-based systems.

3.5 Some project examples

3.5.1 Intelligent user interface for databases (University of Tartu)

The project was aimed at the development of a user interface which enables adaptation to different problem domains and access to different databases. Using minor readjustment, the interface can be tuned to new problem domains. Users enter their query in Estonian and get an answer also in Estonian, in form of a text or as synthesized speech. In the dialogue manager a general model for controlling information dialogues was implemented, which takes into account general regularities of practical dialogues. The Estonian dialogue corpus (Gerassimenko et al. 2008) served as a basic resource for modelling domain-specific conversations. Some other resources for processing of Estonian were integrated into the interface: morphological analysis and generation, spell checking and correction of erroneous forms, automatic recognition of named entities (proper nouns, temporal expressions), and text-to-speech synthesis. Two test applications, user interfaces to a movie schedule database and a dentist database have been developed (www.dialoogid.ee).

Some references: Treumuth (2010); Hennoste et al. (2009); Koit et al. (2009); Koit/Roosmaa/Õim (2009); Koit (2009).

3.5.2 Resources and software for syntactic analysis (University of Tartu)

Syntactic analysis is an important component in different HLT applications including automatic grammar correction, dialogue systems, automatic text summarization, machine translation, etc. The aim of the project was to develop a rule-based syntactic parser for Estonian (based on Constraint Grammar) and its implementations such as a grammar checker for written and spoken language and a prototype for automatic summarization of newspaper texts.

The developed parser gives a shallow surface oriented description of the sentence where every word is annotated with a tag corresponding to its syntactic function (in addition to morphological description). The prototype of grammar checker involves two modules: (1) morphological disambiguation, and (2) syntactic parsing. It gives ca 5% false alarms and misses about 7% of errors.

The prototype of automatic summarization implements different statistic and linguistic methods in order to find the sentences in a text which best represent the content of the analyzed text. The current version is tuned to process news texts on the web (http://lepo.it.da.ut.ee/~kaili/estsum/).

Some references: Lindström/Müürisep (2009); Müürisep/Nigol (2009); Müürisep/Nigol (2008a, b).

3.5.3 Corpus query on the Estonian language website keeleveeb.ee (Filosoft Ltd.)

To enable the use of the Estonian Reference Corpus (www.cl.ut.ee) via www.keeleveeb.ee a convenient query system, allowing the search for lemmas and morphosyntactic categories, was developed. The Estonian Reference Corpus (Kaalep et al. 2010) has been morphologically tagged and disambiguated and clause boundaries have been automatically tagged. The query system to this corpus makes use of all these tags, and in addition, it can be used in conjunction with queries to the dictionaries.

The language portal www.keeleveeb.ee hosts 30 specialised dictionaries, containing over 200,000 concepts. All these dictionaries can be queried simultaneously. In addition, the very same query can also obtain answers from 30 dictionaries hosted elsewhere on the Web, thus linking 60 dictionaries into a single virtual database.

3.5.4 Lexicographer's workbench (Institute of the Estonian Language)

In the project an interactive, web-based working environment for lexicographers was developed. The system called EELex represents a toolset for dictionary management implementing both universal and Estonian-based language resources and linguistic software. EELex makes dictionary work easier and faster, and raises its quality. The dictionaries compiled in or transferred to EELex represent universal reusable language resources with standard XML mark-up, necessary for lexicographers and language technologists. EELex takes care of formatting, punctuation, sorting, referencing, access rights to different sections of the entry and to different working stages etc. Nowadays all dictionaries produced by the Institute of the Estonian Language are compiled using EELex. In addition to the professional system, a public version of the system (http://exsa.eki.ee/) allows dictionary development via Internet for everyone.

Some references: Langemets et al. (2006, 2010).

3.5.5 Research and development of methods for Estonian speech recognition (Institute of Cybernetics at Tallinn University of Technology)

The project is focused on the research, development and testing of methods for Estonian speech recognition and the implementation of speech recognition prototype systems in different domains. The main tasks of the project involved (1) determining optimal basic lexical units for Estonian LVCSR (such as syllables, (pseudo-)morphemes, data-driven units), (2) developing statistical language modelling techniques using the determined lexical units, (3) applying of acoustic model adaptation techniques, (4) developing methods and algorithms for large/unlimited vocabulary speech recognition systems, (5) implementing speech recognition prototype systems.

Main outcomes:

- Complete system for large vocabulary speech recognition of long speech recordings, including speech/non-speech segmentation, speaker diarization, and multi-pass speech recognition, involving unsupervised adaptation techniques.
- Current word error rates: 14.3% for dictated broadcast news, 28.6% for broadcast conversations, 37.1% for conference presentations.

– Web interface (http://bark.phon.ioc.ee/tsab/) for browsing transcribed speech. Supports synchronized listening to speech and viewing its transcriptions, search in transcriptions, viewing related transcripts.
– Speech recognition system for the radiology domain, 9.8% WER with speaker independent models, faster than real time.

References: Alumäe/Kurimo (2010a, b); Ruokolainen/Alumäe/Dobrinkat (2010); Alumäe/Meister (2010); Alumäe, T. (2008).

3.5.6 Centre of Estonian Language Resources

In order to guarantee the availability of the language resources and software prototypes developed in different projects funded by NPELT a project for setting up the Centre of Estonian Language Resources at the University of Tartu was initiated in 2008.

Existing natural language resources can be used by different end users only if they are well documented, archived and publicly accessible. In order to achieve this goal there needs to be an infrastructure to manage and coordinate different activities, from elaborating the corresponding language technology standards to drawing up the contracts/licence agreements necessary for the use of these language resources.

On the European scale, the ESFRI project CLARIN (Common Language Resources and Technology Infrastructure, www.clarin.eu) aims at establishing the infrastructure for documenting, archiving and sharing common language resources. The University of Tartu is the official representative of Estonia among the 36 partners of CLARIN and the Centre of Estonian Language Resources should become a local node of the pan-European infrastructure.

It should guarantee that the existing language resources will not remain only at the disposal of the creators of these resources but will ultimately reach all the interested parties all over Europe, e.g. linguists, teachers, creators of software systems and their applications, civil servants, etc.

In 2010, the Centre of Estonian Language Resources was included in the list of objects of the Estonian Research Infrastructures Roadmap (approved by Government Order No 236 of June 17, 2010) (see https://www.etis.ee/portaal/includes/dokumendid/teekaart.pdf). The Centre is established and will act as a consortium including the University of Tartu, the Institute of Cybernetics at Tallinn University of Technology, and the Institute of the Estonian Language as the main partners.

4. National Programme for Estonian Language Technology (2011-2017)

NPELT 2006-2010 has been definitely successful and has resulted in a remarkable increase of the amount and diversity of language resources and language-specific prototypes. However, the quality and quantity of prototypes and resources is not yet sufficient to enable exploitation of the current technology in end user applications and e-services. Therefore, a follow-up programme was compiled in 2010 and approved by the Minister of Science and Education in January 2011. The follow-up programme is proposed for the period 2011-2017 and supports HLT activities in five action lines:

1. Research and development of language-specific methods and prototypes (speech synthesis and recognition, prosody models for speech synthesis, audiovisual speech synthesis, syntactic analysis adapted to spoken language, semantic analysis, dialogue management, dialogue systems for different domains, analysis of affective speech, tools for translation and terminology management, machine translation, etc.);
2. Development of reusable language resources (text, speech and multimodal corpora, electronic dictionaries and databases, corpus management and access systems, etc.);
3. Support for the Centre of Estonian Language Resources (standardization, licensing, quality control, archiving and documentation of language resources, etc.);
4. Integrated software and application (dialogue systems in specific domains, applications for users with special needs, computer-aided language learning, user interfaces to public services, etc.);
5. Specific projects carried out on demand of the steering committee or to fulfil public needs.

The two first action lines are similar to those of the previous programme; the third one was introduced to support the functioning of the Centre of Estonian Language Resources. Action lines 4 and 5 are new instruments introduced to extend the flexibility of funding schemes. The aim of action line 4 is to promote the use and integration of the existing language resources and prototypes into different applications demonstrating the possibilities of HLT. Action lines 1, 2 and 4 are of bottom-up type, i.e. project applications are proposed by eligible research groups or institutions and the steering committee makes funding decisions based on competition.

Action line 5 is a top-down scheme and is an instrument at the disposal of the steering committee to control the HLT developments more actively. Project tasks and technical requirements are defined by the steering committee or by a public authority and the best bid will be selected.

5. Development of human resources

In Estonia, there exists a critical mass of researchers and engineers working in different HLT areas, and the University of Tartu provides curricula in computational linguistics and in 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 in Linguistics and Language Technology (DSLTT) was launched at Tartu University for 2005 to 2008. The activities of the school have strongly contributed to the effectiveness of many students' doctoral studies; about 10 PhD theses in HLT areas have been prepared and defended with DSLLT support.

In 2009, two new doctoral schools were launched for the period 2009-2015:

– **Doctoral School in Information and Communication Technologies** at Tallinn University of Technology – involves also HLT students from Tartu University;
– **Doctoral School in Linguistics, Philosophy and Semiotics** at Tartu University – involves also students of general linguistics with specialization in computer linguistics.

Estonian language technology researchers are also engaged in the Estonian centre of excellence called **Estonian eXcellence in Computer Science** (EXCS) to be financed over the period 2008-2015. The centre involves the research staff of the Institute of Cybernetics at Tallinn University of Technology, Cybernetica AS, and the University of Tartu representing a major part of the computer science research conducted in Estonia. The general objective of the centre of excellence is to consolidate and advance computer science in 6 areas of recognized strength: programming languages and systems, information security, software engineering, scientific and engineering computing, bioinformatics and human language technology. The specific objectives are to enhance the research potential of the groups by facilitating collaboration, to increase the impact of research results and popularize them in society, and to ensure the sustainability of the groups. This will be achieved by carefully planned coordination and joint actions, targeted at creating a thriving and highly reputed research environment attractive for young researchers.

However, there is a need to improve the study opportunities and attract more students in the speech processing field – currently no systematic teaching is provided in the area of speech analysis, synthesis and recognition.

6. Small and Medium Enterprises (SMEs)

The market situation in Estonia does not attract ICT companies' investments into language-specific HLT developments – there are only ca 1.4 million speakers of Estonian in the world. However, a few small private HLT companies exist:

- **Filosoft** (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 (www.keeleveeb.ee) offering free access to different on-line dictionaries, software and corpora.
- **Keelevara** (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** (www.tilde.ee) is a branch of the Latvian company Tilde (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 mainly on software localisation and translation services.

HLT developments in academic groups typically end up with a prototype which is not yet suitable for an end user application – product development needs a lot of additional work that is beyond the capabilities and interests of an academic researcher. There is a need for an intermediate unit between academy and industry, a development unit which is able to evolve a laboratory prototype into an applicable technology.

In Estonia the competence centres' programme was launched by Enterprise Estonia aiming to bridge the gap between scientific and economic innovation by providing a collective environment for academics, industry and other innovation actors. The compe-

tence centres are established as independent state supported research organizations. Two of eight existing competence centres are potential intermediary units able to carry out industry-oriented applied research and development in HLT field:

– Software Technology and Applications Competence Centre (STACC, established in 2009) – a joint initiative between the University of Tartu and Tallinn Technical University and the leading IT companies and users of Estonian software and knowledge-based technology (including e.g. Cybernetica AS, Regio AS, Webmedia AS, Logica Eesti AS, eHealth Foundation, Skype Technologies OÜ, Swedbank AS). STACC aims to conduct applied research in software technology by working with suppliers and users of technology; among other topics STACC is applying different language technology methods for the analysis of medical text corpora.
– Competence Centre in Electronics, Information and Communication Technologies ELIKO – established in 2004 by Tallinn University of Technology and private companies (including Artec Group OÜ, Apprise OÜ,Cybernetica AS, Modesat Communications, Regio AS, Smartdust Solutions OÜ, Smartimplant OÜ and others). ELIKO is focussing mainly on the development of complex embedded hardware and software systems, but has also done some research in the area of Controlled Natural Language.

It is expected that language resources and language-specific software prototypes developed within NPELT and within the follow-up programme and made available via the Centre of Estonian Language Resources will attract SMEs' and competence centres' interest, leading to the development of end users applications for Estonian market.

7. Summary

The national programme for 2006-2010 has resulted in a remarkable advancement in Estonian HLT. The programme has been successful and has fulfilled most of the expectations. The amount of written and spoken language resources and software prototypes as well as new knowledge and experience created in different projects have strengthened the technological bases for the development of innovative HLT-applications in coming years. To further the HLT progress in Estonia the follow-up programme for 2011-2017 has been launched. It is focused on the development of more advanced software prototypes and new languages resources as well as on the implementation and integration of the software prototypes in public services and commercial applications. The dedicated national initiatives together with international cooperation in EU networks such as CLARIN and META-NET, etc should contribute to the achievement of technological level which allows functioning of Estonian in the modern information society equally with big languages.

8. References

Alumäe, T. (2008): Comparison of different modeling units for language model adaptation for inflected languages. In: Gelbukh, A. (ed.): *Computational Linguistics and Intelligent Text Processing: 9th International Conference, CICLing 2008, Haifa, Israel, February 17-23, 2008. Proceedings*. (= Lecture Notes in Computer Science 4919). Berlin/Heidelberg/New York: Springer, 488-499.

Alumäe, T./Kurimo, M. (2010a): Efficient estimation of maximum entropy language models with N-gram features: an SRILM extension. In: *Proceedings of INTERSPEECH 2010 Spoken Language Processing for All: 26-30 September 2010*. Chiba: ISCA, 1820-1823.

Alumäe, T./Kurimo, M. (2010b). Domain adaptation of maximum entropy language models. In: *48th Annual Meeting of the Association for Computational Linguistics: Proceedings of the Conference, Workshops and Associated Events: Uppsala, Sweden, July 11-16, 2010*. Stroudsburg, PA: Association for Computational Linguistics, 301-306.

Alumäe, T./Meister, E. (2010): Estonian large vocabulary speech recognition system for radiology. In: Skadina, I./Vasiljevs, A. (eds.): *Human Language Technologies: The Baltic perspective. Proceedings of the Fourth International Conference Baltic HLT, Riga, Latvia, October 7-8, 2010*. (= Frontiers in Artificial Intelligence and Applications 219). Amsterdam: IOS Press, 33-38.

Elenius, K./Forsbom, E./Megyesi, B. (2008): Language resources and tools for Swedish: A survey. In: Calzolari, N. et al. (eds.): *Proceedings of the Sixth International Conference on Language Resources and Evaluation: May 26-June 1, 2008, Marrakech, Marocco*. Paris: European Language Resources Association.

Gerassimenko, O./Hennoste, T./Kasterpalu, R./Koit, M./Rääbis, A./Strandson, K./Valdisoo, M./Vutt, E. (2008): Annotated dialogue corpus as a language resource: An overview of the Estonian Dialogue Corpus. In: Shirokov, V. (ed.): *Prikladna lingvistika ta lingvistichni tehnologii: Megaling 2007, Ukraina, september 2007*. Kiev: Dovira, 102-110.

Hennoste, T./Gerassimenko, O./Kasterpalu, R./Koit, M./Rääbis, A./Strandson, K. (2009): Towards an intelligent user interface: Strategies of giving and receiving phone numbers. In: Matoušek, Václav (ed.): *Text, Speech and Dialogue. Proceedings of the 12th International Conference, TSD 2009, Pilsen, Czech Republic, 13-17 September 2009*. (= Lecture Notes in Computer Science 5729). Berlin/Heidelberg/New York: Springer, 347-354.

Kaalep, H.-J./Muischnek, K./Uiboaed, K./Veskis, K. (2010): The Estonian Reference Corpus: Its composition and morphology-aware user interface. In: Skadina, I./Vasiljevs, A. (eds.): *Human Language Technologies: The Baltic perspective. Proceedings of the Fourth International Conference Baltic HLT, Riga, Latvia, October 7-8, 2010*. (= Frontiers in Artificial Intelligence and Applications 219). Amsterdam: IOS Press, 143-146.

Koit, M. (2009): Experiments on automatic recognition of dialogue acts. In: Karpov, A. (ed.): *Proceedings of SPECOM 2009: 13th International Conference Speech and Computer, St. Petersburg, 22-25 June 2009*. St. Petersburg: Institution of the Russian Academy of Sciences/St. Petersburg Institute for Informatics and Automati, 533-538.

Koit, M./Gerassimenko, O./Kasterpalu, R./Rääbis, A./Strandson, K. (2009): Towards computer-human interaction in natural language. In: *International Journal of Computer Applications in Technology*, 34 (4), 291-297.

Koit, M./Roosmaa, T./Õim, H. (2009): Knowledge representation for human-machine interaction. In: Dietz, Jan L.G. (ed.): *Proceedings of the International Conference on Knowledge Engineering and Ontology Development, Madeira (Portugal), 6-8 October 2009*. Setubal: INSTICC, 396-399.

Krauwer, S. (2005): How to survive in a multilingual EU? In: Langemets, M./Penjam, P. (eds.): *Proceedings of The Second Baltic Conference on Human Language Technologies*. Tallinn: Institute of Cybernetics and Institute of the Estonian Language, 61-66.

Krauwer, S. (2006): *Strengthening the smaller languages in Europe. Proceedings of the 5th Slovenian and 1st International Language Technologies Conference, October 9-10, 2006, Ljubljana, Slovenia*. Internet: http://nl.ijs.si/is-ltc06/proc/01_Krauwer.pdf (accessed on 11.06.2007).

Langemets, M./Loopmann, A./Viks, Ü. (2006): The IEL dictionary management system of Estonian. In: de Schryver, G.-M. (ed.): *DWS 2006: Proceedings of the Fourth International Workshop on Dictionary Writing Systems: Pre-EURALEX workshop. Turin, 5th September 2006*. Turin: University of Turin, 11-16.

Langemets, M./Loopmann, A./Viks, Ü. (2010): Dictionary management system for bilingual dictionaries. In: Granger, S./Paquot, M. (eds.): *eLexicography in the 21st century: New challenges, new applications. Proceedings of eLex 2009, Louvain-la-Neuve, 22-24 October 2009*. Louvain-la-Neuve: Presses universitaires de Louvain, Cahiers du CENTAL, 425-430.

Lazzari, G. (2006): *Human Language Technologies for Europe. ITC IRST/TC-Star project report*. Internet: http://tcstar.org/pubblicazioni/D17_HLT_ENG.pdf.

Lindström, L./Müürisep, K. (2009): Parsing corpus of Estonian dialects. In: Bick, E./Hagen, K./Müürisep, K./Trosterud, T. (eds.): *Proceedings of the NODALIDA 2009 workshop Constraint Grammar and robust parsing, Odense, 14.05.2009*. (= NEALT Proceedings Series 8). Tartu: Tartu University Library, 22-29.

Mariani, J. (2009): Research infrastructures for Human Language Technologies: A vision from France. In: *Speech Communication* 51, 569-584.

Meister, E./Vilo, J. (2008): Strengthening the Estonian language technology. In: Calzolari, N. et al. (eds.): *Proceedings of the Sixth International Conference on Language Resources and Evaluation: May 26-June 1, 2008, Marrakech, Marocco*. Paris: European Language Resources Association.

Melero, M./Boleda, G./Cuadros, M./España-Bonet, C./Padró, L./Quixal, M./Rodríguez, C./Saurí, R. (2010): Language technology challenges of a 'small' language (Catalan). In: Calzolari, N./Choukri, K./Maegaard, B./Mariani, J./Odijk, J./Piperidis, S./Rosner, M./Tapias, D. (eds.): *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. Valletta: European Language Resources Association, 925-930.

Müürisep, K./Nigol, H. (2008a): Towards better parsing of spoken Estonian. In: Čermak, F./ Marcinkevičiene, R./Rimkute, E./Zabarskaite, J. (eds.): *Proceedings of the Third Baltic Conference on Human Language Technologies. Kaunas, Lithuania, October 4-5, 2007*. Vilnius: Vytautas Magnus University, Institute of the Lithuanian Language, 205-210.

Müürisep, K./Nigol, H. (2008b). Where do parsing errors come from: The case of spoken Estonian. In: Sojka, P./Horak, A./Kopecek, I./Karel, P. (eds.): *Text, Speech and Dialogue. Proceedings of the 11th International Conference, TSD 2008, Brno, Czech Republic, 8-12 September 2008*. (= Lecture Notes in Computer Science 5246). Berlin/Heidelberg/New York: Springer-Verlag, 161-168.

Müürisep, K./Nigol, H. (2009): Shallow parsing of transcribed speech of Estonian and disfluency detection. In: Vetulani, Z./Uszkoreit, H. (eds.): *Human Language Technology. Challenges of information society. Third Language and Technology Conference, LTC 2007, Poznań, Poland, October 5-7, 2007*. Berlin/Heidelberg/New York: Springer-Verlag, 165-177.

Odijk, J. (2010): The CLARIN-NL Project. In: Calzolari, N./Choukri, K./Maegaard, B./Mariani, J./Odijk, J./Piperidis, S./Rosner, M./Tapias, D. (eds.): *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. Valletta: European Language Resources Association, 48-53.

Ruokolainen, T./Alumäe, T./Dobrinkat, M. (2010): Using dependency grammar features in whole sentence maximum entropy language models for speech recognition. In: Skadina, I./ Vasiljevs, A. (eds.): *Human Language Technologies: The Baltic perspective. Proceedings of the Fourth International Conference Baltic HLT, Riga, Latvia, October 7-8, 2010*. (= Frontiers in Artificial Intelligence and Applications 219). Amsterdam: IOS Press,73-79.

Spyns, P./D'Halleweyn, E. (2010): Flemish-Dutch HLT policy: Evolving to new forms of collaboration. In: Calzolari, N./Choukri, K./Maegaard, B./Mariani, J./Odijk, J./Piperidis, S./Rosner, M./Tapias, D. (eds.): *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. Valletta: European Language Resources Association, 2855-2862.

Treumuth, M. (2010): A Framework for Asynchronous Dialogue Systems. In: Skadina, I./Vasiljevs, A. (eds.): *Human Language Technologies: The Baltic perspective. Proceedings of the Fourth International Conference Baltic HLT, Riga, Latvia, October 7-8, 2010*. (= Frontiers in Artificial Intelligence and Applications 219). Amsterdam: IOS Press, 107-114.