Appendix 1
CONFIRMED
By the Minister of Education and Research
25 January 2011
directive no 71

National programme

Estonian Language Technology 2011-2017

Introduction

Language technology is an interdisciplinary field integrating information technology and linguistics which is concerned with developing language software and language resources necessary for the computational processing of human language.

Language software involves methods for processing language materials, algorithms and computer programmes and is the basis for language technology application systems. It is useful to distinguish between **speech technology** (e.g. speech recognition and speech synthesis) and the **technology for processing written texts** (e.g. morphological, syntactic and semantic analysis), i.e. language technology in its more restricted sense. **Language resources** are electronic databases that are used to develop language software: corpora (collections of speech signals and texts), electronic dictionaries and databases.

The development of **language technology** support of the Estonian language is necessary for the preservation of the linguistic and cultural identity of self in the multilingual environment of the EU and for communicating on equal basis with technologically advanced languages in the information society.

The programme is designed as a follow-up programme for the "National Programme for Estonian Language Technology (2006−2010)" which made a considerable contribution to the development of language technology in Estonia; however, the language resources and the prototypes of the programme are not yet sufficient enough to ensure the functioning of the Estonian language in the e-environment. The follow-up programme differs from the previous programme in that in addition to the development of software prototypes and language resources, considerable attention is paid to creating language technology applications and making accessible the existing resources and software as well as those to be created.

The general directions in the field of language technology software and its applications and in the field of language resources are the following:

- in speech technology: 1) the level of speech recognition allows its use outside laboratory conditions, 2) the quality of speech synthesis approaches that of natural language;
- in language technology: 1) automatic analysis and synthesis on every language level (morphology, syntax, semantics, pragmatics), 2) the availability of application systems (data mining, machine translation, resources for translators and interpreters, workbench for lexicographers and terminologists, etc.);
- in integrated applications: the integration of different language and speech technology resources for developing the applications necessary for end-users;

- the field of language technology resources requires the availability of corpora and dictionaries which meet the needs of software development and application projects.

The ministry responsible for the national programme is the Ministry of Education and Research; as a potential partner from the public sector, cooperation is carried out, first and foremost, with the Department of State Information Systems of the Ministry of Economic Affairs and Communications.
The main partners as the main participants of the programme are the University of Tartu, the Institute of Cybernetics at Tallinn University of Technology, and the Institute of the Estonian Language, including also other research and development institutions and information technology companies.

Background

One of the fundamental values of the European Union is linguistic and cultural diversity and in order to preserve this diversity and to develop dialogues between different cultures, various pieces of EU legislation have been introduced and numerous projects to develop and promote language technologies have been funded[1]. Although all of the official languages of the European Union hold equal status, bigger languages dominate the domains of business and technology development, since for languages with smaller numbers of speakers language technology development is not cost-effective. Thus, it is pointed out in the report charting the language technology needs of the EU[2] that from the perspective of language technology as a business field one can distinguish between primary, secondary, and tertiary languages. It is clear that the safeguarding of the equality of languages in the management of EU takes its toll on the tax payers – with 23 official languages it is necessary to guarantee translation between 506 language pairs. Around 1% of the EU budget is spent on translation and interpreting, for example in 2007 it was around 1 billion euros; this figure is predicted to rise by 5% annually[3].
Language technology development is equally expensive for every language; it does not make a dramatic difference whether it is done for English or Estonian. It is enormously expensive to raise the language technology standards of all the official languages of the EU to the levels of English and it is

[1] COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS. Multilingualism: an asset for Europe and a shared commitment. Brussels, 18.9.2008, COM(2008) 566.

[2] G. Lazzari. Human Language Technologies for Europe. ITC IRST/TC-Star project report, 2006. http://www.european-journalists.eu/Human%20Language%20TechHnolLogies%20for%20Europe%20-%20TC-STAR.pdf

[3] http://www.euractiv.com/en/culture/eu-translation-policy-stay/article-170516

unavoidable that both the European Commission as well as Member States themselves should invest in it (in accordance with the principle of subsidiarity).

The state of language technology in Estonia in 2010

The tasks of the "Development Strategy of the Estonian Language 2004−2010" in developing the language technology support of Estonian have been accomplished: the Estonian language belongs to the top 50 technologically advanced languages of the world. The following national programmes have helped to achieve this: "Estonian Language and Cultural Memory (2004–2008)" which funded language technology work during the years 2004–2005, "National Programme for Estonian Language Technology (2006−2010)", and the doctoral school "Linguistics and Language Technology" active at the University of Tartu during 2005–2008. Prior to 2004, the development of language technology support of Estonian was supported by IT and CS Education Development Centre and the national programme "Estonian Language and National Culture" (1999-2003).

However, the size and diversity of different language resources and the number and quality of the developed software prototypes are not yet on the level that would enable the widespread application of language technology in the e-environment.

The current situation, problems and needs in higher education

There is a critical mass of specialists to develop language resources, language software and its applications in Estonia. New specialists can be trained at the University of Tartu at the specialities of Estonian and Finno-Ugric Linguistics (specialising in computational linguistics) and Information Technology (specialising in language technology at the MA level). In 2009, two new doctoral schools have been launched where the doctoral students of computational linguistics and language technology participate: the National Doctoral School in Information and Communication Technologies (including language technologists) and the Graduate School of Linguistics, Philosophy and Semiotics which also incorporates computational linguists via general linguistics.

The problem lies in training speech technology specialists – the BA, MA and PhD curricula of the Faculty of Information Technology at Tallinn University of Technology include only a few elective courses in signal and speech processing and prerequisite courses providing the necessary mathematical basis, but there is no systematic advanced study in the analysis, synthesis and recognition of speech signals.

In order to achieve the objectives set forth in the Development Plan of the Estonian Language (2011–2017) – that language technology support for the Estonian language will be on an equal level with that

of other languages in countries with advanced language technology (e.g. the Nordic countries) – and in order to develop the information and communication services used by language technology, it is necessary to maintain and ensure the training of a sufficient number of researchers and developers in the field of language technology in both written and spoken language. It would be expedient to arrange for the training of specialists to take place in cooperation between the University of Tartu and Tallinn University of Technology.

The current situation, problems and needs in research

There are three main centres for language technology research and development in Estonia:

(1) Research Group of Computational Linguistics at the University of Tartu, with the following key research areas: morphological, syntactic, semantic and pragmatic analysis; spoken language and models of dialogue, machine translation and the creation of the relevant language resources (written language corpora, semantic databases, corpora of dialogues and spontaneous speech, parallel corpora). Language software and resources have also been developed in other research groups: bioinformatics research group (fuzzy data mining, text algorithms) and phonetics research group (the phonetic database of spontaneous speech).

(2) The research and development activities of the Research Group of Language Technology at the Institute of the Estonian Language include resources for lexicographers and Estonian lexicon databases, methods of text-speech synthesis, including the analysis and synthesis of emotional speech. There are 16 people working in the field of language technology at the Institute of the Estonian Language, 12 of them full-time and 4 part-time.

(3) The key research areas of the Laboratory of Phonetics and Speech Technology of the Institute of Cybernetics at Tallinn University of Technology are experimental-phonetical research and the development of different speech corpora (read, spontaneous and dialogue speech, radio news and talk shows, foreign-accented speech, academic lectures, etc.). During the years 2006–2010, 2 senior researchers and 1 researcher/doctoral student were involved in this research area on a long-term basis at the institute; 1 researcher and 2 doctoral students on a short-term basis; 1 MA student, 2 BA students and 2 other employees with authorisation agreements.

These three research centres have made a substantial contribution to the completion of NPELT (2006–2010) objectives; a few language technology projects have also been carried out at the University of Tallinn (Estonian Interlanguage Corpus) and Estonian Literary Museum (compilation of an electronic base dictionary of Estonian idiomatic expressions).

The state-financed research projects at the University of Tartu, the Institute of the Estonian Language, and the Institute of Cybernetics at Tallinn University of Technology contribute significantly to the

development of language-specific theoretical models and knowledge important for language technology development. The "Centre of Excellence in Computer Science" (2008–2015) has been launched where language technologists from the University of Tartu and the Institute of Cybernetics at Tallinn University of Technology also participate. Estonian language technologists regularly attend international conferences and participate in various EU networks (CLARIN, META-NET, HEXA-NORD, etc.).

The infrastructure necessary for research is of relatively high quality in each research group. The Centre of Estonian Language Resources, launched as part of the Estonian Research Infrastructure Roadmap, helps to improve local and international cooperation between research groups; the Centre of Language Resources will be responsible for making the language resources and technologies developed by the partners accessible to researchers.

The main problem lies in the relative shortage of researchers and developers-programmers within the field of speech technology; at least as many as 2 to 3 researchers and 3 to 4 engineers/programmers are needed in addition. There is a lack of programmers qualified to fulfil the tasks of electronic lexicography, for example the Institute of the Estonian Language needs at least 2 full-time programmers, while one researcher with the relevant programming know-how will retire soon. In order to maintain the quality of research it is necessary to continuously invest in modernising the infrastructure of the research groups and in research training.

The current situation, problems and needs in entrepreneurship

There are only a few companies active in the field of language technology in Estonia:

- Filosoft (http://www.filosoft.ee, founded in 1993) – the main products are the speller and the hyphenator for Estonian, a thesaurus for different operational systems (MS Windows, Unix, Mac OS) and software packages (MS Office, OpenOffice, Lotus Notes). Filosoft is also responsible for administrating the Internet portal Keeleveeb (http://www.keeleveeb.ee), through which a number of dictionaries and text corpora can be accessed for free.
- Many Estonian dictionaries, translation and specialised dictionaries, and the databases of names have been made available through the web portal of Keelevara (http://www.keelevara.ee, founded in 2004).
- Tilde Eesti (http://www.tilde.ee) is the Estonian branch of the Latvian company Tilde founded in 1991. Tilde is the biggest and best known company in the Baltic States in the field of language technology. The main products of the company are the word processing resources (localised fonts, spellers, grammar checkers), electronic dictionaries, resources for translators and interpreters, etc. adapted to the languages of Eastern Europe. Tilde Estonia offers services mainly for the localisation and translation of software.

- Software Technology and Applications Competence Center (http://www.stacc.ee) is an organisation funded through the Enterprise Estonia Competence Centre programme (2009–2015) which aims, among other things, to develop language technology methods for the analysis of medical texts.

The reason for the relatively low interest of Estonian ICT companies in developing language technology applications lies in the fact that language technology development is based on long-term research work and requires large-scale investments which are not financially cost-effective considering the size of the Estonian language market. At the same time, the development of language technology applications and the development of Estonian language e-environment using these applications, is first and foremost, the task of companies and not research groups. The accessibility of language resources and software prototypes to be developed through the national programme are an important prerequisite for the creation of applications supporting the Estonian language that are oriented towards the end-user.

Estonian language technology in comparison with world-wide and EU trends

The policies of the European Union in treasuring and developing the linguistic and cultural diversity are presented in the various documents of the Commission of the European Communities, e.g. "A New Framework Strategy for Multilingualism"[4] (2005) and "Multilingualism: An Asset for Europe and a Shared Commitment"[5] (2008). In addition to other measures (promoting language teaching, subtitling media productions, creating information content for multilingual multimedia, etc.), considerable attention is also paid in these documents to the necessity of language technology development. In addition to the measures planned by the Commission, the Member States are encouraged to support the development of language technology and to establish national plans that would help to give structure, coherence and direction to actions to promote multilingualism.

The EU information society technologies programme supports research work concerned with the crossing of language barriers with the help of new information and communication technologies:

- resources to increase the efficiency of translators and interpreters (translation memories, online dictionaries and thesauri), semi and fully automatic translation systems;
- speech recognition and synthesis, dialogue systems.

The communication from the Commission (2008) states that "faced with the globalising online economy and ever-increasing information in all imaginable languages, it is important that citizens access and use information and services across national and language barriers, through the internet and mobile devices. Information and communication technologies (ICT) need to be language-aware and promote content creation in multiple languages."

European information society framework documents "i2010 – A European Information Society for Growth and Employment" (2005)[6] and "A Digital Agenda for Europe"[7] (2010) aim to build a fully inclusive European information area, for which measures are planned to develop the services of

[4] OPINION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS: A new framework strategy for multilingualism, 22.11.2005, COM(2005) 596.

[5] COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS. Multilingualism: an asset for Europe and a shared commitment. Brussels, 18.9.2008, COM(2008) 566.

[6] COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS: i2010 – A European Information Society for Growth and Employment. Brussels, 01.06.2005, COM(2005) 229.

[7] COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS: A Digital Agenda for Europe. Brussels, 26.8.2010, COM(2010) 245.

information and communication technologies and to increase the necessary investments. The building of an inclusive information area requires the development of language technology which enables to cross language barriers and provides access to multilingual information and services and encourages citizens to participate in the e-environment.

As stated by "A Digital Agenda for Europe" the Commission shall present an in-depth research and innovation strategy as part of the flagship initiative "Innovation Union" in 2010 in order to implement the Europe 2020 Strategy.[8] According to the strategy, in order to establish its leadership in ICT[9], Europe must raise, monitor, and centre its investments in order to maintain competitiveness in this area and it must continue to invest in high risk research and development, including basic research covering many fields. The Commission is planning the following measures for this:

- it guarantees sufficient financial support for joint ICT R&D infrastructures and innovation clusters, continues to develop the e-infrastructure and sets the EU's strategy for cloud data processing for, first and foremost, government and research;
- it collaborates with stakeholders in order to build a new generation in web-based applications and services, including multilingual information content and network services, supporting the standards and open platforms through programmes funded by the EU.

Member States should:

- double by 2020 the national gross expenditure in ICT research and development from 5.5 billion euros to 11 billion euros (including EU programmes) in such a way that the expenditure in the private sector would also increase from 35 billion euros to 70 billion euros;
- participate in large-scale pilot projects in order to test and develop innovative and collaborative solutions in areas offering public interest which are financed by CIP (competitiveness and innovation framework programme).

A number of pan-European networks have been launched in the European Union, e.g. CLARIN, META-NET, etc., the aim of which is to build infrastructures to ensure the accessibility of language resources and technologies and to develop language technology for all EU languages. There are a number of areas in the EU Seventh Framework Programme as a part of which projects involving language technology are also financed, e.g. cognitive systems, communication and robotics; digital content and libraries; etc. At the same time, the last EU framework programmes (FP6 and FP7) have neither directly touched upon the subject matter of developing language technology resources and base software, nor the activities directed at languages with a less developed technology support.

[8] The 2009 Report on R&D in ICT in the European Union. http://ftp.jrc.es/EURdoc/JRC49951.pdf

[9] A Strategy for ICT R&D and Innovation in Europe: Raising the Game, COM(2009) 116.

In a number of EU Member States, national programmes have been launched during the last few decades to develop language technology, for example:

- In France, in 1994 *Francil*-network and in 2002 *Techno-Langue French national programme (http://www.technolangue.net)*;
- In the Netherlands, the language technology coordinated development began in 1999 and in 2005 the Dutch-Flemish joint programme STEVIN (2005-2011) (http://taalunieversum.org/taal/technologie/stevin/) was launched in order to develop the technology support for the Dutch language;
- In 2000 the Nordic network for language technology (http://cst.dk/nordoknet) and a language technology research project (2000-2004) for Danish, Icelandic, Norwegian, Swedish and Finnish was launched;
- In the Finnish technology agency TEKES the following programmes include language technology: USIX (1999-2002) and FENIX (2003-2007) (within this programme, the sub-programme PUMS is directly aimed at language technology);
- Sweden's innovation agency VINNOVA financed a language technology programme during the years 2001-2004.

Language technology has mostly been developed for English; the French, German, Spanish, Italian, Japanese and Chinese languages also have a strong technological support. At a relatively good level is the technology support for the Dutch, Danish, Swedish, Finnish, Norwegian, Czech, Hungarian, Polish, Portuguese, Greece and Slovenian languages. Technologically weaker are the Bulgarian, Irish, Latvian, Lithuanian, Romanian and Slovakian languages.

The measures adopted in Estonia for developing the Estonian language and its language technology support (Development Strategy of the Estonian Language (2004–2010), the NPELT (2006–2010)) agree in every respect with the EU trends highlighted above. NPELT has made a considerable contribution to the development of language technology – the size and diversity of language resources and the number and quality of prototypes have increased considerably, which allows the Estonian language to be placed among other languages with a good language technology support.

As is well known, not many countries coordinate the development of language technology at a national level and for this reason the NPELT has drawn a lot of attention at a number of international conferences (LREC 2008, NODALIDA 2009, BalticHLT 2010, EFNIL 2010).

The objectives and results of the "National Programme for Estonian Language Technology (2006–2010)"

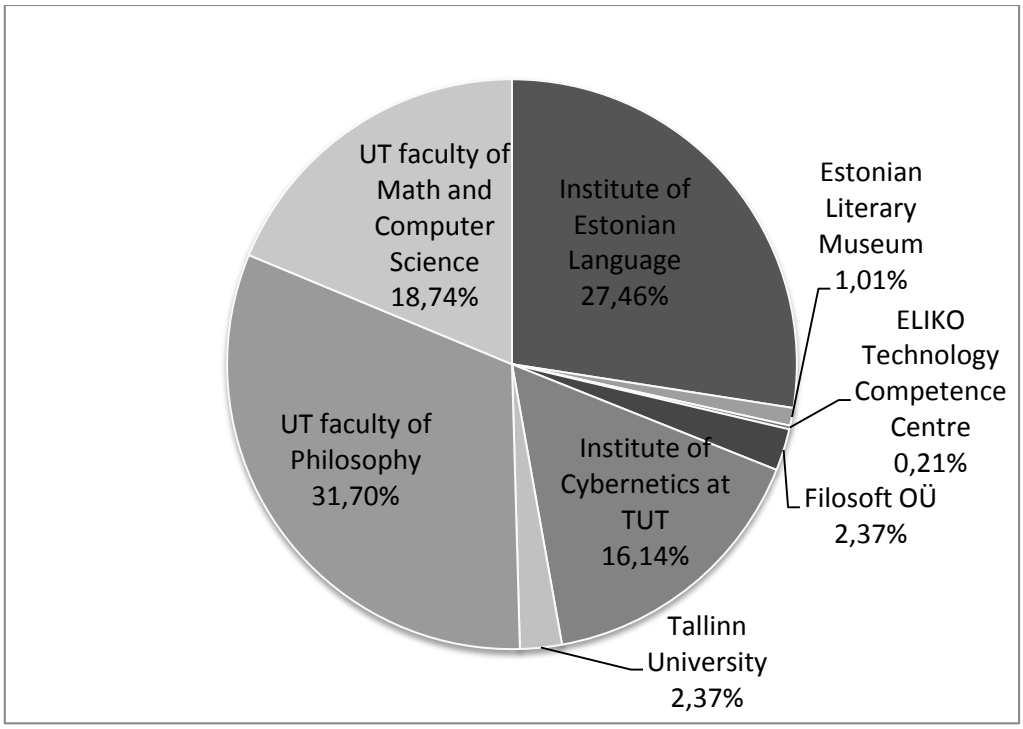
The main objective of the "National Programme for Estonian Language Technology (2006–2010)" (NPELT) was to advance the language technology support for the Estonian language to the level that would enable the Estonian language to function successfully in today's information technology environment. NPELT has financed research and development in language technology from the building of resources to the building of prototypes for language technology applications.

All of the projects financed through NPELT are aimed at achieving the main objective of the programme and the programme on the whole has been successful. In total 22 projects were financed during the years 2006–2010; 18 of these were concerned with research into language technology methods and the creation of software prototypes, 14 with the compilation of different language resources and the development of their user interface and one project was concerned with the planning of infrastructure necessary for the management of language resources and software.

Table 1. The number of projects and funding per year, the number of follow-up projects and new projects and the recalculation of the programme financing into thousand euros is given separately in the parentheses.

	2006	**2007**	**2008**	**2009**	**2010**
The number of financed projects	**18**	**20** (18+2)	**23** (20+3)	**23** (15+8)	**24** (22+2)
Programme financing, million kroons (thousand euros)	**7,3** (466,5)	**7,1** (455,7)	**13,4** (856,4)	**12,9** (842)	**11,8** (765,3)

The distribution of funding according to the institutions:



The steering committee highlights some of the projects that characterise most vividly the achievements of NPELT 2006–2010.

The success stories of NPELT 2006-2010

Estonian dialogue on the computer (University of Tartu, project leader M.Koit)

Web-based software has been developed during the projects *Estonian Information Dialogue on the Computer* and *Interactive Information Retrieval System for Estonian* that can be adapted to different fields and integrated with different databases. The retrieval system can be attuned to new fields and integrated with databases by introducing minimal modifications, giving thus the user the possibility to access databases in Estonian and to receive as output adequate and true information. Software for the automatic processing of Estonian are integrated in the intelligent retrieval system: morphological analysis and synthesis, the checking of spelling and the correction of incorrect forms, the identification of temporal expressions and proper names, text-speech synthesis. The retrieval system has been tested with cinema and dental care information databases. See www.dialoogid.ee.

Language software based on syntactic analysis and the language resources necessary for its development (University of Tartu, project leader T.Roosmaa)

The summarizer EstSum of Estonian texts is aimed at generating summaries of online news articles and is currently a prototype version. Automatic summary generation of texts is a process during which a new, abridged version of the already existing texts is created which contains only the information necessary for the user.

Naturally the computer does not understand the content of the text to be processed nor the user's expectations vis-à-vis the summary. Therefore, a combination of different statistical and linguistic methods are used in order to find sentences within the text that are considered to be the most comprehensive and rich in information and the user is presented with a collection of such sentences as the summary.

Although the resources and methods used for automatic summary generation are generally independent of language, the same resources developed for one language cannot be used to generate summaries in another language. The morphological and syntactic characteristics of the language need to be taken into account, as well as the standardised expressions suitable for summaries, and the lexical differences.

The programme is available for testing on the following web-page: http://lepo.it.da.ut.ee/~kaili/estsum/

Corpus query in the Estonian language website keeleveeb.ee (Filosoft, project leader H.-J.Kaalep)

30 specialised dictionaries, containing over 200 000 concepts, are made available for free on www.keeleveeb.ee. Some key aspects:

1. Simultaneous queries can be run with all of the specialised dictionaries; the very same query can also obtain answers from 30 dictionaries outside www.keeleveeb.ee. This means that one is able to search for a word or a term in as many as 60 dictionaries simultaneously.

2. Among other things, 14 subject-specific dictionaries for primary schools which were compiled at the University of Tartu on the order of the Ministry of Education and Research and published in 2005 are made available. The dictionaries list the Estonian terms together with the explanation of the term and its Russian translation. For many subjects, the dictionaries are the first dictionaries in Estonian within the specific field.

3. Queries to the Estonian Reference Corpus (http://www.cl.ut.ee/korpused/segakorpus/, 250 million words) can be run together with queries to dictionaries. The Estonian Reference Corpus has been morphologically tagged and disambiguated, indexed by word form, lemma, and grammatical information during its integration with keeleveeb. A number of usage examples can be obtained this way in addition to the dictionary definitions; the former may specify the information provided by the latter.

Lexicographer's workbench (Institute of the Estonian Language, project leader Ü.Viks)

As a result of the project, web-based lexicographer's workbench (EELex) has been set up that brings together the software and language resources necessary for a lexicographer, supports group work and offers support in Estonian. On the basis of the professional lexicographer's workbench a public consumer version of EELex has been created (http://exsa.eki.ee/) which enables the users to compile their own dictionaries. EELex makes dictionary compilation and editing easier, faster, and raises its quality.

What sets EELex apart from other world-wide electronic dictionary systems is the existence of Estonian language support (integrated automatic morphology, database of the Estonian-X dictionary), its enormous flexibility as pertains to the structure of the dictionary, rich choice in editing tools and free access. All of the dictionaries published by the Institute of the Estonian Language are compiled in the EELex system. Cooperation has been initiated within Estonia, as well as outside, with publishers and other institutions where dictionaries are compiled or prepared for use with new applications (e.g. telephone dictionaries).

Research and development of methods for Estonian speech recognition (Institute of Cybernetics at Tallinn University of Technology, project leader T.Alumäe)

The project is concerned with research, development and testing of methods for Estonian speech recognition and the implementation of speech recognition prototype systems for the Estonian language. The project included the following subfields: (1) determining the optimal recognition units for Estonian speech recognition (diphones, syllables, pseudo-morphemes, etc.), (2) morpho-syntactic language models based on different recognition systems and the problems related to adapting statistical language models, (3) modelling semantic connections in statistical language models, (4) modelling Estonian quantity degrees on the basis of the relationships between syllable lengths, (5) technological solutions for creating speech recognition systems with limited and unlimited vocabulary.

The outcomes of the project are numerous prototypes:

- Forced alignment – software that automatically segments Estonian speech into words and sounds;
  http://www.phon.ioc.ee/dokuwiki/doku.php?id=projects:tuvastus:est-align.et
- Complete system of automatic segmentation and transcription of speech recordings (e.g. radio shows, conference presentations and lectures);
- Web interface for browsing, indexing and searching transcriptions of recorded speech; http://bark.phon.ioc.ee/tsab/

- Speech recognition system for the radiology domain (in cooperation with Cybemetica AS).

Estonian Language Technology Roadmap (2011–2017)

In 2005 Estonian Language Technology Roadmap (2004–2011) was drawn describing the state of language technology in 2004 and envisioning the development of language resources and prototypes until 2011. Majority of the envisioned results have been achieved and/or the respective research and development is in progress; it transpired that the achievement of some objectives is considerably more time-consuming than initially envisioned due to the complicated nature of the task, for example, building the prototypes for the automatic recognition of dialogues, the semantics-pragmatics interface, the audiovisual speech synthesis, and the dialogue system enabling the use of spoken language.

Estonian Language Technology Roadmap (2011–2017) establishes the status of Estonian language technology resources and prototypes as of the end of 2010 and envisages the developments for the next seven years. In order to map the existing language resources, prototypes and the needs a questionnaire has been conducted among language technology researchers-developers and potential end-users, the results of which form the basis for drawing up a new roadmap. The roadmap shall be confirmed as an appendix of the programme during 2011.

Legislation and long-term development plans underlying the programme

Development Plan of the Estonian Language 2011–2017

The Development Plan of the Estonian Language 2011–2017 (Estonian Language Foundation, Government of the Republic order 26.11.2010 no. 451) presents in the chapter titled “Language Technology Support of the Estonian Language” the analysis of the situation, objectives for the development of language technology support of the Estonian language, its impact on society and its indicators. The Development Plan of the Estonian Language sets the following objective:

- the level of language technology support of the Estonian language is on par with the languages of language-technologically advanced countries (e.g. the Nordic countries) in areas which are required by the developments and applications of software aimed at the Estonian language.

The impact on society upon completing the development plan consists of two points:

- programs that adequately process language data are extensively used in the society
- the Estonian language is recognized as one of the languages with an advanced language technology, Estonia is one of the countries whose language technology is the most advanced

In order to develop the language technology support of the Estonian language it is necessary:

- to support the participation of Estonian language technologists in the international division of labour, creation of open-source applications, the use of standard protocols and standardising our own resources and solutions
- to build a central depository to manage language resources and reusable software

The indicators for the completion of the development plan are concerned with the opportunities to use natural speech synthesis, speech recognition created for limited domains, man-machine dialogue system, and machine-translation prototype.

Estonian Information Society Strategy 2013

The strategy states that in the development of the information society the sustainability of the Estonian language and culture is ensured. The challenges listed in the strategy include access to computers and the Internet, information society infrastructure, the use of the Internet in households and enterprises, and competitiveness of the ICT sector.

The information society development envisages an all-inclusive information society that is constantly developing and raising the living standards and a competitive Estonian economy where rational ICT solutions are used in order to achieve greater productivity and employment rate.

The present programme is in agreement with the strategy and its aims.

Estonian Research and Development and Innovation Strategy "Knowledge-Based Estonia 2007-2013"

The Estonian research and development and innovation strategy stipulates that national research and development programmes will be launched for solving socio-economic problems and achieving the objectives in socio-economic sectors that are important to every resident of Estonia, as for instance information society and its relation to ensuring and promoting the sustainability of the Estonian language.

Links to other national and international programs

Estonian Research Infrastructure Roadmap

The Estonian Research Infrastructure Roadmap, approved by the Government of the Republic, lists the Centre of Estonian Language Resources (CELR) as one of the nationally important objects. The centre shall act as a dislocated infrastructure as prescribed by the consortium agreement signed by the University of Tartu, Tallinn University of Technology, and the Institute of the Estonian Language. CELR shall be the CLARIN-ERIC (European Research Infrastructure Consortium) centre in Estonia.

In developing the language technology support it may prove to be necessary to cooperate with the other research infrastructure roadmap object "Estonian E-Repository and Conservation of Collections".

National programmes

NPELT (2006–2010) has evolved from the national programme "Estonian Language and National Memory (2004–2008)"; the follow-up programme of the latter – "Estonian Language and Cultural Memory (2009–2013)" (ELCM) – includes three directions: 1. the Estonian language, 2. cultural memory, 3. raising the quality of publications in humanities. The sub-activity 1.2. of the first direction "the Estonian language" stipulates the following "Regulating, digitalising, and publishing of linguistic databases" if the databases built for language technology solutions are not supported and the supporting activities shall not duplicate the activities of the language technology national programme – in order to guarantee the latter, cooperation shall be ensured with the steering committee of ELCM and it will also be checked in the future that the ELCM endeavours shall not overlap with the endeavours of the 2nd sub-objective of the present programme. The national programme for the support of Estonian terminology (2008–2012) includes among other activities the building of an IT environment supporting the development of terminology and the compilation of a public terminology database at the Institute of the Estonian Language. The text of the programme suggests the standards to be coordinated with the steering committee of NPELT; the current situation could favour the unification of standards with the Centre of Language Resources among whose consortium partners shall also be the Institute of the Estonian Language.

CLARIN

CLARIN (www.clarin.eu) is an ESFRI roadmap object, the goal of which is to create a pan-European infrastructure of the existing language resources and language technologies with the aim of making these resources and technologies accessible to all researchers, especially researchers from the humanities and social sciences. The goal of CLARIN is to offer high-quality services that would cross language and field specific boundaries and that would help the preservation and active use of the multilingual and multicultural richness of Europe. CLARIN tries to overcome the current situation which is characterised by the fragmentation and dis-coordination of the existing language data and resources by offering supranational and supra-field web services. CLARIN is based on national infrastructures/centres of different countries that are able to manage national language resources and technological solutions.

Uniform standards, quality requirements, etc. are established through CLARIN and agreements are made as to the principles of the access rights of resources. Thus all the resources of different

languages included in partner centres are made easily accessible to European researchers through the CLARIN centres. In order for the Estonian researchers to benefit from this wealth of pan-European language resources and technologies and in order to add Estonian language resources to this infrastructure, the Centre of Estonian Language Resources shall be created.

The sub-objectives and expected results of the programme

The programme is divided into 5 sub-objectives.

1. Research and development projects for building software prototypes

Research and development projects for building software prototypes in the following fields (the list is incomplete):

- speech recognition:
    - research and development of language and application-specific recognition modules (acoustic models, language models)
    - applications in different fields (the automatic transcription of spontaneous and dialogue speech, radio and TV shows, the dialogue systems of limited fields, speech recognition in channels of communication with limited frequency band)
    - automatic identification of spoken language
- speech synthesis:
    - speech synthesis interfaces and applications in different fields (generation of digital books, audio systems for people with a visual disability, the voicing of subtitles in digital television network, interface for using text-speech synthesis in different systems and application programs, automatic generation of new synthesis voices from random speech corpora)
    - integration of the prosodic models of speech synthesis with other levels of language (syntax, semantics, pragmatics) and extralinguistic factors (emotions)
    - audiovisual speech synthesis (the so-called talking head) models
- syntactic/semantic/pragmatic analysis-synthesis:
    - syntactic parser which takes into account the idiosyncrasies of spoken language
    - syntactic synthesis on the basis of semantic representation
    - 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)

- analysis-synthesis of cohesive texts (including dialogues) in speech and writing:
    - 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)
    - automatic recognition/synthesis of dialogue acts
    - automatic recognition/synthesis of dialogue strategies
    - prototypes for dialogue systems and interface (in specific fields)
    - (semi-)automatic transcription system for spoken language
    - automatic recognition of extralinguistic communication signals (voice quality, emotions, laughter, etc.)
- text processing devices:
    - automatic text type identifier
    - workbench for translators and interpreters
    - workbench for terminologists
    - modules for modifying the workbench for lexicographers
    - indexation system for Estonian web documents enabling the identification of authorship and plagiarism
- machine translation:
    - application of devices for processing Estonian in machine translation system

Projects shall be applied for within an open competition, the maximum duration of a project is up to 4 years; competitions are held annually.

2. Projects for building language resources

Projects for building and developing re-usable language resources in the following fields (the list is incomplete):

- text corpora:
    - corpus of Internet language
    - learner corpus
    - balanced corpus of literary language
    - corpus of Estonian scientific texts

- syntactic/semantic resources:
    - treebanks
    - semantic database
    - frame lexicon
    - domain-specific ontological databases
- lexicographic resources:
    - lexico-grammatical database
    - dictionary databases
- machine translation resources:
    - parallel corpora for the needs of statistical machine translation (EU documents, the fields of international communication most important for Estonian researchers)
- spoken language resources:
    - corpora for different types of natural speech (spontaneous, emotional, speech with a foreign accent, corpora for domain and text type specific speech)
    - dialogue corpora
- multimodal resources:
    - audio-video corpora
    - sign language corpus
    - databases for speech production (glottography, palatography)
- development, standardisation, and free access of electronic dictionaries and ontological databases
- development of tagging, search and management systems of corpora

Projects shall be applied for within an open competition, the maximum duration of a project is up to 4 years; competitions are held annually.

3. Centre of Estonian Language Resources

To create a central depository for managing language resources and software – the Centre of Estonian Language Resources (CELR)

The Centre of Estonian Language Resources is an infrastructure created on the basis of the consortium agreement signed by the partners through which Estonian language resources are made accessible to the interested parties. In addition to collecting and archiving the existing language resources as well as those to be created as a result of NPELT 2006–2010, the centre shall launch a system for the recognition of collected language resources and for the training of potential users.

Natural language resources can be used by different applicants and interested people only when the existing language resources are duly documented and archived and made publicly accessible. In order to support these activities which may seem at times unnecessary for the creators of language resources it is necessary that a certain infrastructure exists that would arrange and coordinate work in this field in Estonia. The centre's field of activity shall be from the creation/fixation of language technology standards to the compilation of legal contracts/licences necessary for the use of language resources.
The Centre of Estonian Language Resources to be created tries to do its best by using the know-how involved in the projects of CLARIN in that the existing language resources of Estonia would not only be known by the creators and builders, but that they would reach all possible interested parties, such as linguists, teachers, creators of software systems and applications, civil servants, etc.
The present project may finance the following CELR activities: fixation of standards, compilation and signing of licence agreements, inspection of access rights, quality inspection and documentation of archived resources, creation of user environments, PR and training, etc.
Projects shall be applied for according to the activity plans of the centre.

4. Integrated language software and its applications

Integrated language software and its applications:

- man-machine-dialogue systems in restricted domains, e.g. in directory enquiries
- technical aids for people with special needs
- computer and/or Internet-based language learning
- interface for public services

Projects shall be applied for within an open competition, the maximum duration of a project is up to 2 years. The competition shall not be held annually, the date for the announcement of the competition shall be decided by the steering committee of the programme. A prerequisite for the project application is the involvement of partners from the public or private sector, it is required that the partner(s) make a contribution (financial or intellectual) that can be measured realistically.

5. Development projects to be ordered

Development projects ordered on the proposal of the steering committee or a public sector company.

The task and technical requirements of the project shall be compiled by the customer; projects shall be applied for within an open competition, the maximum duration of the project is up to 2 years; the

competition shall not be held annually, the date for the announcement of the competition shall be decided by the steering committee of the programme.

Development activities

Development activities shall take place within all sub-objectives, but first and foremost within sub-objectives 4 and 5. The Ministry of Economic Affairs and Communications, but also other ministries and state institutions are envisioned as the main public sector partners. By making accessible the resources and software prototypes to be created within the programme, business enterprises are stimulated to use language technology software and to create new innovative e-services in the interest of both the public sector as well as in business applications.

Licensing the use of created resources and software

Principles:

- language resources and software prototypes created within the programme are intellectual property,
    - for which the choice of means of protection are guided by the aim to promote the introduction of the created property, Open Access licensing is preferred, proceeding from non-profit intentions and avoiding the transfer of exclusive proprietary rights;
    - the use of which for different objectives (public use, research, business application) is regulated by different types of licences which may be both free and not free and which may depend on other legal regulations, for instance protection of personal data.
- the Centre of Estonian Language Resources shall manage, make accessible and license the created resources and software.

All of the outcomes of the programme projects shall be deposited as the final versions of the project (together with the documentation) at CELR and a framework agreement shall be signed for the use of the project outcomes. In very exceptional circumstances, when the project's final outcome cannot be deposited at CELR, the project participant shall ensure the preservation of the final version of the project outcome and its use on its own servers and data media.

Programme management

For the substantive management of the programme the steering committee of the programme shall be formed. The steering committee of the programme shall be formed from language technology experts, representatives of the Ministry of Education and Research, the Ministry of Economic Affairs and Communications and the general research public.

Programme steering committee

The task of the steering committee is to analyse the development of language technology in Estonia and in the world. The steering committee shall compile the application forms of the programme and decide separately each year for which sub-objectives it shall open the call of applications. The steering committee shall distribute the means allocated for the sub-objectives of the present programme within an open competition or purposefully for the work of the Centre of Estonian Language Resources on the basis of the agenda, bearing in mind that the proportions between sub-objectives shall be confirmed separately each year proceeding from the number and quality of the submitted project applications.

The steering committee shall be responsible for the completion of the programme's objectives and shall ensure that the means of the programme are used purposefully. In order to achieve this, the steering committee shall make broad-based use of the help of experts. The programme steering committee shall form working groups to accomplish specific tasks.

Programme administration and coordination

The steering committee shall delegate the administration of the programme at the earliest opportunity to the Centre of Estonian Language Resources, who shall find within an open competition the programme coordinator. The Ministry of Education and Research shall sign an agreement for the administration of the programme with the institution curating the Centre of Estonian Language Resources who in turn shall sign a contract of employment with the programme coordinator. The expenses of the programme coordinator and for programme administration shall be covered from the programme means.

The programme coordinator shall participate in the meetings of the steering committee, but he or she shall not have the right to vote. The programme coordinator shall be responsible for arranging the programme competition, the purposeful use of the means and for compiling reports. The programme coordinator shall prepare the contracts to be signed with programme participants. The programme coordinator has the right to make proposals to the steering committee as to the hiring of additional labour within the budget. The coordinator shall present the substantive reports about the completion of

the programme to the steering committee for confirmation and financial reports to the Ministry of Education and Research.

Promotion and public relations

The programme coordinator shall engage in promoting the programme and in its public relations.

Implementation of the programme

The programme is designed to last for seven years. During the first years, the sub-objectives 1 and 2 have priority, during the last three years the sub-objectives 4 and 5 shall gain more importance.

Financial needs of the programme

The financial needs of the programme in euros.

	2011	2012	2013	2014	2015	2016	2017	Total cost
1. Software prototypes	443866	499150	499150	499150	499150	499150	499150	**3438766**
2. Language resources	229635	226567	226567	226567	226567	226567	226567	**1589037**
3. Centre of Estonian Language Resources	76502	76502	99702	108650	117597	127823	134214	**740990**
4. Integrated language software	0	72731	69664	110886	161697	209950	225768	**850696**
5. Development projects to be ordered	0	14380	68705	105454	161697	209950	225768	**785954**
Programme administration	15339	18215	25245	27802	35151	38027	40264	**200043**
Total	**765342**	**907545**	**989033**	**1078509**	**1201859**	**1311467**	**1351731**	**7605486**

The sums given in the table are indicative, the actual financing depends on the state budget. The dynamics that during the first years the sub-objectives 1 and 2 have the priority and that during the final three years the sub-objectives 4 and 5 shall gain in importance has been taken into account in compiling the table.

The desired situation upon the implementation of the programme and the critical factors from the perspective of effectiveness

The aim of the programme is to ensure a situation where the Estonian language technology support is on equal level with that of other languages in countries with advanced language technology (e.g. the Nordic countries) in directions that require the software developments and applications oriented to the Estonian language. It has also been noted for languages with a more advanced language technology that many language technology tasks do not have as easy a solution as previously estimated, especially the creation of dialogue systems for machine translation and spoken communication have proven to be much more difficult tasks than previously assumed. Therefore, the probability of achieving the completion of the programme objectives by the year 2017 is 70–80%.

Impacts of the programme:

- programs that process language materials adequately have found a broad use in society; it is possible to use natural speech synthesis, speech recognition created for specific domains, widespread technical aids for text editors, man-machine-dialogue systems and machine translation prototypes;
- the Estonian language shall be recognised as a language with advanced language technology, Estonia belongs to the countries with most advanced language technology.

Critical factors

The successful completion of the present national programme depends on the management of the programme, responsible work by parties (both the main participants as well as the business sector), training and existence of specialists and the optimal financing of the programme. In case the programme is underfunded, a choice has to be made and some of the objectives cannot be completed. Since language technology solutions presume the completion of consecutive objectives, underfunding may mean that language technological resources are created, but language software and language technology solution prototypes are not. At the same time the funding of the programme that is larger than optimal does not entail that the programme is completed quicker because the number of people working in the field of language technology is limited in Estonia. An important risk factor in completing the programme is the incomplete application of the training programme for necessary specialists, including those with a PhD degree, at the universities. When the salaries in the private sector reach the levels of the European Union and the programme is underfunded, people will leave the academia for the private sector and the objectives of the programme will not be completed. The funding of the programme has to be adapted to the general rise in salaries in comparable jobs in the private sector.

The objectives of the programme cannot be completed without the modernisation of Estonian language technology infrastructure, i.e. without the modernisation and unification of NPELT participants' hardware and software and without a common information technological platform and standards. In order to achieve this, the launching of the Centre of Estonian Language Resources makes an important contribution. A possible competition between the programme participants and the faltering of cooperation between the other participants of the programme and the private sector are potential risks. In order to minimise the risk a widespread cooperation with local as well as international partners is necessary.

The existing IT solutions that have a closed source code or are patented and that not all users can afford but which at the same time present an argument to turn down the financing of duplicate language technology solutions, present a risk for the programme to a certain degree. In order to minimise the programme risks it is important to develop modern licensing policy and to ensure the freeware of language technology solutions that play a key role in the development of the Estonian language.

Another set of risks includes a potential change in language attitudes which would bring about a widespread use of software in English and which would make the creation of Estonian IT environment useless. Involved in this risk is the widespread commercial onset of software in English. Creating Estonian versions of these software would help to minimise the risk.

In addition, the fact that language technology may be fetishised and people will expect more from the programme than it can offer, neglecting thus the everyday (practical) language planning, poses a potential risk for the completion of the programme.

Indrek Reimand
Head of Research Policy Department