Human Language Technologies – The Baltic Perspective
A. Tavast et al. (Eds.)

doi:10.3233/978-1-61499-133-5-273

Strategic Importance of Language Technology in Estonia

Kadri VIDER [1], Krista LIIN and Neeme KAHUSK
Institute of Computer Science, University of Tartu

Abstract. This paper gives an overview of the strategies and national programmes related to the development of language technology in Estonia. It also describes briefly the international initiatives aiming to build infrastructures for gathering and sharing language resources that Estonian linguists and language technology researchers participate in.

Keywords. Estonian, national programme, research infrastructure, CLARIN, META-NET

Introduction

Language technology belongs to a strategic key technology — info and communication technology — of the Estonian RD&I Strategy 2007–2013 “Knowledge-based Estonia” [1]. That strategy stipulates that Estonian researchers have to participate in the international division of labour, create open code application whenever possible, and use standard protocols and other standardization measures for language resources and technological solutions too.

Language technology is also a key technology in the “Development Strategy of the Estonian language” [2] where the goal of LT is to develop the technology support for the Estonian language to the level that would allow functioning of Estonian language in the modern information society.

In these 2 strategies the main reasons are formulated, why is it strategically important for Estonia to deal with HLT development.

1. National Programmes

1.1. Establishing the Strategy for Estonian LT

The beginning of a systematic approach to planning Estonian language technology research and development belongs to the year 1995, when Estonia joined the EU COPERNICUS programme. The Estonian Informatics Centre was established in 1997 with the aim to develop national info systems. One part of its activities was the Estonian Language

[1] Corresponding Author: Kadri Vider, Institute of Computer Science, University of Tartu, Liivi 2, 50409 Tartu, Estonia; E-mail: kadri.vider@ut.ee

Technology Programme (1998–2000) that supported the creation of basic LT resources (such as OCR, morphological analyzer, speech synthesis or thesauri). [3]

The first Development Plan for Estonian Language Technology [4] was formed in 1999 under the Estonian Language Technology Programme. As a next, more concrete step in forming the strategy, the Roadmap of Estonian Language Technology for 2004–2011 [5] was compiled in 2004 as a result of the eVikings II project [6]. It gave the baseline as the resources that had been developed by 2004 and presented a timeline of the future developments in Estonian LT in three major action lines: spoken language technology, written language technology and language resources (including the infrastructure for collecting and managing these resources).

In several EU-countries diverse national level initiatives are undertaken in order to facilitate and coordinate research and development of HLT for national languages.

In Estonia, the National Programme for Estonian Language Technology (NPELT) was launched in 2006 [7]. NPELT is financed from state budget, in 2006–2010 ca 54 MEEK (3,4 M€) in total. For period 2011–2017 it is planned to fund different projects and activities with ca 1 M€ every year. NPELT management is carried out by a Steering Committee including HLT experts and representatives of the ministries and ICT industry.

All results of the NPELT projects (language resources and software prototypes) are released as public domain. The Center of Estonian Language Resources has a duty to deposit all such resources and tools for preservation and long term access.

1.2. First Phase of the National Programme

NPELT in its first phase (2006–2010) [8] funded HLT–related R&D activities including the creation of reusable language resources and development of essential linguistic software (up to the working prototypes) as well as bringing the relevant language technology infrastructure up to date.

33 projects, targeting these sub-goals, were funded in the frame of the first phase. Projects covered wide range of topics: speech synthesis and recognition, compilation of digital language corpora, Estonian emotional speech corpora, database of an Estonian–X dictionary, lexicographer’s workbench, machine translation, information dialogue, lexical resources and tools, morphosyntactic and semantic analysis, language learning, and language software for the Web.

Projects falling under the sub–goal of creation of reusable language resources gave the most results — the existing Estonian language resources were significantly improved on, both in size, annotation and standardization, and new text and speech resources allowing for deeper linguistic information (such as deeper syntax or semantics in corpora, emotions in speech corpora) were developed [9]. The chosen projects have more or less followed the topics and goals set in the Estonian HLT Roadmap for 2004–2011 [5]. As a result, the state of Estonian speech and text resources is on the same level as many European languages, such as Danish or Finnish [9].

Most noticeable results in creating software belong to the field of speech technology — considerable improvement on the speech synthesis (project “Corpus-based speech synthesis for Estonian” [10]) and the development of speech recognition for Estonian (project “Research and development of methods for Estonian speech recognition” [11]). Both the development of the tools and gathering of necessary corpora were supported by the programme, not only are the results available, but also integrated into applications that have reached end-users such as speech synthesis for Windows, speech synthesis tool for the blind or speech recognition application on smartphones.

Other examples of the LT tools resulting from the first phase of NPELT are Estonian–English machine translation that became available before Google included Estonian in their machine translation system, or the lexicographer's workbench which aids in creating machine-readable, standard–adhering dictionaries.

The third sub-goal — bringing the relevant language technology infrastructure up to date was implemented by the project preparing the creation of Center of Estonian Language Resources.

1.3. Evaluation of the First Phase of NPELT (2006–2010)

The first phase of NPELT has created favourable conditions for HLT development in Estonia. The amount of re–usable language resources and software prototypes as well as new knowledge and experience created within the NPELT will serve as the technological bases for the development of innovative HLT–applications in coming years [12].

Ministry of Education and Research had the Archimedes Foundation Research Council do a target evaluation of the first phase of NPELT. According to the Steering Committee that evaluated the performance of the projects, at least 84% of the projects had satisfactory results. The target evaluation report [13] states that 79% of the projects managed to partly or completely fulfil their goals. Although the main goal of NPELT was for the results of the projects to be used in the modern information society, in this phase the research and development projects were funded up to the creation of prototypes. This discrepancy between the goals of the projects and the funded activities led to the failure of covering/making the important step from a preliminary prototype to an actual product ready for the end-user. Another uncovered aspect is the need for continued development and support of a marketed product based on that software prototype.

There are also some aspects from outside the scope of the programme that limit the usability of its results — of which the issues related to intellectual property rights are the most eminent. Issues such as the use of copyrighted resources in the development of the product, the use of freeware modules in commercial applications or the obscurity of property rights in case one of the parties involved in the development of the product is a commercial enterprise often make it difficult to make use of the language resources and software resulting from the NPELT programme.

1.4. Second Phase of the National Programme (2011–2017)

The second phase of the NPELT started in 2011 [14] and focus on the implementation and integration of the existing resources and software prototypes in public services. This phase pay more attention on NLP applications and software prototypes, and to dissemination of available language resources. In order to achieve the aims of the programme, in second phase five type of actions will be funded:

1. R&D to create new software;
2. Projects for new language resources;
3. Center of Estonian Language Resources;
4. Projects for integrating language-specific software into other applications;
5. Specially aimed projects (on the order of Steering Committee)

There have been 3 calls for projects in the years 2011 and 2012, in the frame of which 9 projects to create new software and 9 projects for new language resources have been funded. 2 new projects for integrating language-specific software into other applications started in 2012 and call for applications in specially aimed projects is being prepared by an expert work group.

Intellectual property created according to this phase of NPELT will be covered with different types of user licenses for different purposes. Results of projects as software and resources will be disseminated via Center of Estonian Language Resources.

2. Building up Large-scale Research Infrastructure for Estonian LT

For strategic planning it is necessary to map the existing language resources and situation of different European languages in digital age — this is also important goal of META-NET [15] and task of META-NORD [16] project, where Language White Papers for Baltic and Nordic languages [17] describing the LT landscape for each of those languages were prepared.

In order to increase the visibility and usability of language resources and facilitate their integration in the modern information society, there are several actions taken in the META-NORD in addition to describing the national LT situation [18]. Language resources from the participating countries are brought to the same standards, which makes it easier to use them separately or combine them into multilingual resources.

In collaboration with the pan-European META-NET, common licences are being developed to suit the needs of language resources, e.g. solutions for combined resources consisting of several parts (such as monolingual parts or independent corpora) or for resources that have many property holders (such as reference corpora collected in the course of decades). The solutions are designed to work throughout Europe, facilitating the use of LR across borders.

The goal of increasing international visibility of language resources is reached through making the resources available through the META-SHARE network [19], which is composed of network nodes sustained by project partners that function both as metadata editors and as search and browse tools for the resources. The uniform metadata schema allows for consistent resource descriptions and is linked to the standardised data categories, which makes it easier to map it to other metadata formats. The metadata is to be automatically harvested and updated between the network nodes.

16 Estonian LT resources have already been shared through the META-SHARE node builded up at University of Tartu [20]. These include mostly written resources — different text corpora (8), dictionaries (3), lexical resources (2) that already have well spread standards for representation, but also a speech corpus and text analysis tools (2). Other language resources are in the process of being described and formatted according to the META-SHARE format.

The work on systematically connecting the Estonian language resources to international networks began in participation as one of the 36 European partners in the CLARIN (Common Language Resources and Technology Infrastructure) initiative [21]. CLARIN aims to facilitate the use of language resources for researchers of other fields — humanities and social sciences, but also for linguists and society members. That means both creating a stable integrated infrastructure and providing support for users with little technical knowledge about the usage of language resources or tools.

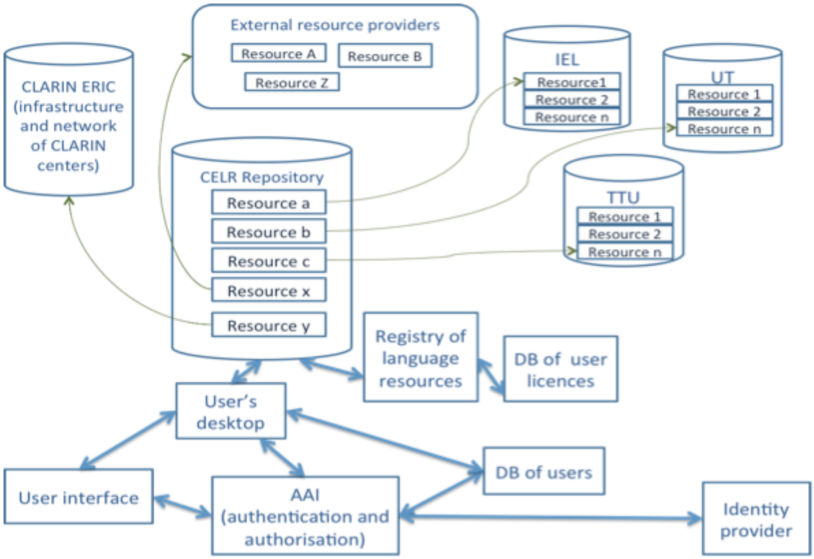

Figure 1. CELR as infrastructure: components and structural schema.

As one part of the preparatory phase of CLARIN in 2008–2011, 32 Estonian language resources were described according to and included in the Virtual Language Observatory [22] — a repository that gathers language resources from both the CLARIN partners as well as harvesting metadata from several other LR inventories. In addition to preparing the language resources with regard to interoperability and sharing, the work in building the infrastructure included surveys of both the existing language resources and the needs of users from different research fields, work on technical interoperability, agreement on data and interoperability standards etc.

2.1. Center of Estonian Language Resources

Estonia has set up the Center of Estonian Language Resources (CELR) as a consortium of 3 institutions at the national level on 2nd of December 2011. This consortium of University of Tartu, the Institute of Cybernetics at Tallinn University of Technology and the Institute of the Estonian Language will perform as organisational framework for coordinating and implementing the obligations of Estonia as the member in CLARIN ERIC (European Research Infrastructure Consortium). The CELR is an infrastructure of national importance in the Estonian Research Infrastructures Roadmap [23] and will provide access to language resources and technologies for all researchers.

To achieve this, the existing digital language resources will be interconnected and supplemented by language technology tools as an environment with web-based access and services which will use the archived/stored data. In the starting stage (2012–2015), CELR will build up an infrastructure of central data register and service servers, user authentication and authorisation systems, system for gathering standardised, well-documented and evaluated collections of data (Figure 1). The auditing and authorization of the users will take place on different levels. Restricted access and the use of resources will be regulated by user licences either between institutions or between an institution and a user.

As CLARIN ERIC by its Statute strongly recommends Open Source and Open Access principles, user conditions and licences based on Creative Commons and General Public Licence are favoured.

Acknowledgements

The work within the project META-NORD has received funding from the ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, Grant agreement no 270899.

CLARIN preparatory phase project received funding from the Preparatory phase for the projects in the 2006 ESFRI Roadmap subprogramme of the EC Seventh Framework Programme, Grant agreement no 212230.

In startup-phase (2012–2015), Center of Estonian Language Resources will be funded from EU Structural Funds from measure "Modernising of research infrastructure of national priority" [24]. CELR is also partly supported from state budget of Estonia as a project EKT10 from National Programme of Estonian Language Technology.

References

[1] Estonian Research and Development and Innovation Strategy 2007–2013 "Knowledge-Based Estonia". http://www.hm.ee/index.php?03242.

[2] Estonian Language Council. Development Strategy of the Estonian language 2004–2010. http://eki.ee/keelenoukogu/strat_en.pdf.

[3] Heiki-Jaan Kaalep and Haldur Õim. Eesti keeletehnoloogia sihtprogrammi arengust. In *Infotehnoloogia haldusjuhtimises. Aastaraamat 1999.*, pages 47–51, Tallinn, 1999. http://www.riso.ee/et/pub/1999it/24.htm.

[4] Heiki-Jaan Kaalep, Einar Meister, and Haldur Õim. Eesti keeletehnoloogia arenduskava. http://www.eki.ee/keeletehnoloogia/tutvustus/arenduskava.html.

[5] Deliverable D3.3: Estonian Language Technology Roadmap, 2004. http://ev2.ioc.ee/dels/wp3/D3.3.doc.

[6] EU FP5 project "eVikings II: Establishment of the Virtual Centre of Excellence for IST RTD in Estonia"(2002–2005),.

[7] Einar Meister, Tiit Roosmaa, and Jaak Vilo. Estonian Language Technology Anno 2009. *Proceedings of the NODALIDA 2009 workshop Nordic Perspectives on the CLARIN Infrastructure of Common Language Resources : May 14, 2009, Odense, Denmark*, (5):21–26, 2009. http://hdl.handle.net/10062/9669.

[8] National Programme for Estonian Language Technology. Riiklik programm "Eesti keele keeletehnoloogiline tugi (2006–2010)". http://www.keeletehnoloogia.ee/ekktt-1.

[9] Krista Liin, Kadri Muischnek, Kaili Müürisep, and Kadri Vider. Eesti keel digiajastul — The Estonian Language in the Digital Age. In Hans Uszkoreit and Georg Rehm, editors, *META-NET White Paper Series*. Springer, 2012. Available online at http://www.meta-net.eu/whitepapers/volumes/estonian.

[10] Project "Corpus-based speech synthesis for Estonian" website. http://www.keeletehnoloogia.ee/projects-1/corpus-based-speech-synthesis-for-estonian.

[11] Project "Research and development of methods for Estonian speech recognition" website. http://www.keeletehnoloogia.ee/-projects-1/research-and-development-of-methods-for-estonian.

[12] Einar Meister, Jaak Vilo, and Neeme Kahusk. National Programme for Estonian Language Technology: A Pre-final Summary. In Inguna Skadiņa and Andrejs Vasiljevs, editors, *Human Language Technologies: The Baltic Perspective. Proceedings of the Fourth International Conference Baltic HLT 2010*, pages 11–14. IOS Press, 2010.

[13] Riikliku programmi "Eesti keele keeletehnoloogiline tugi (2006–2010)" sihtevalveerimine. Evalveerimisraport. Technical report, Archimedes Foundation, 2011. http://www.keeletehnoloogia.ee/varia/ekktt-sihtevalveerimise-raport.

[14] National Programme for Estonian Language Technology. Riiklik programm "Eesti keeletehnoloogia (2011–2017)". http://www.keeletehnoloogia.ee.

[15] META-NET website. http://www.meta-net.eu/.

[16] META-NORD project website. http://www.meta-nord.eu/.

[17] META-NET White Paper Series. http://www.meta-net.eu/whitepapers.

[18] Andrejs Vasiljevs, Markus Forsberg, Tatiana Gornostay, Dorte Haltrup Hansen, Kristín Jóhannsdóttir, Gunn Lyse, Krister Lindén, Lene Offersgaard, Sussi Olsen, Bolette Pedersen, Eiríkur Rögnvaldsson, Inguna Skadiņa, Koenraad De Smedt, Ville Oksanen, and Roberts Rozis. Creation of an Open Shared Language Resource Repository in the Nordic and Baltic Countries. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/744_Paper.pdf.

[19] META-SHARE network node at ELDA. http://metashare.elda.org.

[20] META-SHARE network node at the University of Tartu. http://metashare.ut.ee.

[21] CLARIN (Common Language Resources and Technology Infrastructure) website. http://www.clarin.eu.

[22] CLARIN Virtual Language Observatory. http://www.clarin.eu/vlo/.

[23] Estonian Research Infrastructures Roadmap, 2010. https://www.etis.ee/portal/Portaal/includes/dokumendid/Teekaart.pdf.

[24] European Union Structural Assistance to Estonia. Measures in the period of 2007–2013. http://www.struktuurifondid.ee/meetmete-nimekiri-inglise-keeles/.