

NEALT PROCEEDINGS SERIES

VOL. #

*Nordic Perspectives on the CLARIN Infrastructure
of Common Language Resources
– Workshop Proceedings*

May 14, 2009

Odense, Denmark

Nodalida 2009

Editors

Rickard Domeij
Kimmo Koskenniemi
Steven Krauwer
Bente Maegaard
Eiríkur Rögnvaldsson
Koenraad de Smedt

NORTHERN EUROPEAN ASSOCIATION FOR LANGUAGE TECHNOLOGY

Nordic Perspectives on the CLARIN Infrastructure of Language Resources
– Workshop Proceedings
NEALT Proceedings Series, Vol. #

© 2009 The editors and contributors.

ISSN 1736-6305

Published by

Northern European Association for Language
Technology (NEALT)
<http://omilia.uio.no/nealt>

Electronically published at

Tartu University Library (Estonia)
<http://dspace.utlib.ee/dspace/handle/10062/4116>

Volume Editors

Rickard Domeij
Kimmo Koskenniemi
Steven Krauwer
Bente Maegaard
Eiríkur Rögnvaldsson
Koenraad de Smedt

Series Editor-in-Chief

Mare Koit

Series Editorial Board

Lars Ahrenberg
Koenraad De Smedt
Kristiina Jokinen
Joakim Nivre
Patrizia Paggio
Vytautas Rudzionis

Preface

The Nordic countries have a long tradition of cooperating within many areas, including politics, education and science. Many languages are closely related and sometimes also the same language is spoken over national boundaries (for example Sámi and Swedish). Language technology is relatively well developed in these countries, but much more is needed to build the infrastructure needed for advanced R&D, and to secure the languages of the region for the future. The CLARIN project is an initiative on the European level to meet those challenges by making language resources and technology available and usable.

In recent years, new regions around the Baltic have become parts of the Nordic area. With increased cooperation, coordination and consolidation of common strengths, the Nordic/Baltic countries could strengthen their work in language technology infrastructure and their contribution to CLARIN.

The main topic of the workshop is Nordic strengths and opportunities of cooperation within the NEALT Geographic Region in constructing an infrastructure of common language resources in connection to the European CLARIN (Common Language Resources and Technology Infrastructure) initiative.

The purpose is to find ways of cooperating that will strengthen the contribution of the associated countries to the development of an infrastructure of common language resources within the CLARIN initiative. In the workshop, participants in CLARIN from the NEALT-associated countries will be given the chance to present their national projects and discuss possible ways of cooperating, sharing resources, coordinating activities, consider new projects and such. Opportunities and proposals for closer cooperation and coordination will be presented and discussed at the workshop.

CLARIN participants in the NEALT-associated countries were invited to present their national work from the perspective of possible cooperation between groups and projects in the different countries: Denmark, Estonia, Finland, Iceland, Latvia, Lithuania, Norway and Sweden. The workshop is intended for participants having an interest in developing the language resources and technology in the NEALT Geographic Region for languages spoken in that region. There will be an opportunity for each such country to present an overview of the status of their national language resource infrastructure.

Eight abstracts from each of the above mentioned countries were submitted for review by the editors (who did not review contributions from their own country). All of them were accepted. We, the editors, want to thank the authors for their contributions. We look forward to a promising workshop in Odense where they will be presented and discussed.

Rickard Domeij, Kimmo Koskenniemi, Steven Krauwer,
Bente Maegaard, Eiríkur Rögnvaldsson & Koenraad De Smedt

*Nordic Perspectives on the CLARIN
Infrastructure
of Common Language Resources
– Workshop Proceedings*

May 14, 2009

Odense, Denmark

Nodalida 2009

Workshop website:

<https://kitwiki.csc.fi/twiki/bin/view/Nealt/SigInfraNodalida2009WorkshopCall>

Conference website:

<http://beta.visl.sdu.dk/nodalida2009/>

Organizers

Rickard Domeij, the Language Council of Sweden, chairman of the Nordic language councils' working group on language technology and chairman of the NEALT [SigInfra](#)

Kimmo Koskenniemi, Prof. of language technology at the University of Helsinki, Department of General Linguistics, Executive Board member of CLARIN

Steven Krauwer, Utrecht University, CLARIN Coordinator

Bente Maegaard, Head of Centre for Language Technology, University of Copenhagen, Denmark. Executive Board member of CLARIN

Eiríkur Rögnvaldsson, Prof. of Icelandic Language, University of Iceland

Koenraad De Smedt, Prof. of Computational Linguistics, University of Bergen; national coordinator of CLARIN in Norway

Swedish CLARIN Activities

Maia Andréasson

Lars Borin

Markus Forsberg

Språkbanken

Dept of Swedish Language

University of Gothenburg

first.last@svenska.gu.se

Jonas Beskow, Rolf Carlson

Jens Edlund, Kjell Elenius

Kahl Hellmer, David House

Centre for Speech Technology

School of Computer Science

and Communication

KTH

(rolf,kjell,davidh)@speech.kth.se

Magnus Merkel

NLP Lab

Dept of Computer Science

Linköping University

mme@ida.liu.se

Eva Forsbom, Beáta Megyesi

Language Technology Unit

Dept of Linguistics and Philology

Uppsala University

first.last@lingfil.uu.se

Anders Eriksson

Phonetics Unit

Dept of Philosophy, Linguistics

and Theory of Science

University of Gothenburg

anders.eriksson@ling.gu.se

Sven Strömqvist

Centre for Languages and Literature

Lund University

sven.stromqvist@ling.lu.se

Abstract

Although Sweden has yet to allocate funds specifically intended for CLARIN activities, there are some ongoing activities which are directly relevant to CLARIN, and which are explicitly linked to CLARIN. These activities have been funded by the Committee for Research Infrastructures and its subcommittee DISC (Database Infrastructure Committee) of the Swedish Research Council.

1 Introduction

CLARIN <<http://www.clarin.eu>> has two partners (Centre for Speech Technology, KTH and the Humanities Lab, Lund University) and a considerable number of members in Sweden, including the sites of the authors of this document.

However, the Swedish Research Council has decided not to allocate national funds for Swedish involvement in the ongoing preparatory phase of CLARIN, which means that any participation by Swedish members beyond that which is covered by EC funding to the two Swedish CLARIN partners must be covered by funds obtained elsewhere.

On the other hand, the Swedish Research Council has increased available funding for research

infrastructure *in general*, and in fact Swedish CLARIN members have been able to secure project funding for some CLARIN-related activities in this way from the Committee for Research Infrastructures and its subcommittee DISC (Database Infrastructure Committee) of the Swedish Research Council.

CLARIN-related work in Sweden has been considerably aided by the fact that the Swedish language technology community is close-knit – with well-functioning channels and fora of communication and collaboration – and united in its recognition that the realization of the kind of infrastructure that CLARIN engagement requires is a costly endeavor which must be a collective undertaking involving the whole community.

In the next section we describe some of the ongoing CLARIN-related activities in Sweden, for which we have been able to secure funding by the Swedish Research Council.

2 Some CLARIN-related activities in Sweden

2.1 An infrastructure for Swedish language technology

In 2007, the Research Infrastructure Committee of the Swedish Research Council awarded a two-year

planning grant to a national Swedish consortium in language technology, with 7 partner institutions:

- University of Gothenburg (coordinating partner)
- Chalmers University of Technology
- KTH (Royal Institute of Technology)
- Linköping University
- Lund University
- The Swedish Language Council
- Uppsala University

The planning grant was awarded for a proposal entitled *An infrastructure for Swedish language technology*, with the aim of preparing a project proposal or project proposals for creating an integrated basic Swedish language technology research infrastructure, consisting of

1. a Swedish national corpus (*Svensk nationell korpus* – SNK);
2. a Basic LAnguage Resource Kit (BLARK) for Swedish.

The practical planning work has been carried out by two working groups, with researchers from Gothenburg and Linköping responsible primarily for the work on SNK, and researchers from KTH and Uppsala having worked mainly on the Swedish BLARK. The two groups have interacted constantly throughout the course of the work, both in physical meetings and by means of electronic communication, e.g., project reports and other documents have been collectively prepared using a project wiki.

The main tasks of the working groups have been:

- to make an inventory of and collect information about existing resources, their character, quality, and not least, availability for research and other purposes;
- to make a survey of the needs of the research community and industry;
- to collect information about similar initiatives – completed, ongoing and planned – in other countries, especially in Europe;
- on the basis of this information, to formulate a concrete funding proposal to VR/KFI, comprising a description of the SNK and the Swedish BLARK, together with an outline work plan and budget for creating the resources.

A funding proposal for an SNK/BLARK combination was submitted to VR/KFI in October 2008. The proposal is now being reviewed by international experts. The amount of funding needed for realizing the SNK and Swedish BLARK in parallel is estimated at 130 million SEK over 7 years. However, it is pointed out in the proposal, that pursuing the two separately would cost on the order of 50 million SEK more, i.e., there is considerable synergy in the proposal.

No doubt in large part as a result of the work in this planning project, the Swedish Research Council has listed language technology as one of a number of national research infrastructure areas of highest priority in its *Roadmap to research infrastructure*. This spring, a call will be issued for proposals by national consortia in exactly those areas. Thus, it seems there is a good chance that the two years of dedicated work laid down in this project might pay off.

2.2 Safeguarding the future of Språkbanken

Språkbanken (the Swedish Language Bank; <<http://spraakbanken.gu.se>>) at the University of Gothenburg provides an online service to the research community since 1975, whereby language resources (corpora and lexicons) are made available to the research community and the public. The resources are available free of charge on the internet through a number of search interfaces. Språkbanken possesses a unique combination of competences in the areas of Swedish text corpora, parallel text corpora, Swedish computational lexicons, and LT tools for the processing, annotation and presentation of text corpora, coupled with the kind of stable organization required for sustained large-scale corpus processing and presentation.

Språkbanken's resources are widely used for research and teaching, but also for other related purposes (for checking what is possible or good Swedish, as a reference in popular writings about language usage, etc.). In particular, a good number of PhD theses in Sweden and Finland have used Språkbanken as a data source.

Språkbanken has grown organically over the four decades of its existence. Many of the presently available corpora have been collected on Språkbanken's own initiative, and this is ongoing work; e.g., about 15–20 million words of press text are added annually. However, some of the corpora are the result of independent research

projects conducted by the NLP research group at Gothenburg or by groups at other Swedish universities. In principle, the same situation obtains for the lexicon resources. Tools for browsing and searching resources have been developed in concert with the creation of the resources themselves. This means that resources are stored in Språkbanken in several different formats, with varying amounts of added information. The use of different formats implies that idiosyncratic tools are required for browsing and searching each resource. A number of language technology tools are used with the resources, which have been developed or adapted in various research projects in the department. There are also tools that have been developed in collaboration with other groups, e.g. morphological processors for modern Swedish and Old Swedish which are being developed jointly with the Language Technology research group at Chalmers University of Technology. The conditions under which such research endeavors are undertaken have not in general been conducive to standardization and wider integration of these tools.

Generally, the kinds of research questions which can be addressed using a large text material such as that found in Språkbanken are heavily dependent on three characteristics of the material and the infrastructure in which it is embedded: (1) the character of the material itself (its representativity w.r.t. the language variety under investigation); (2) the annotations, markup and metadata that the material is provided with (and, more generally, which annotations, etc., are [formally] allowed by a given framework); (3) the level of access to the material, viz. (3a) inspection (search and presentation) access only: (3a1) restricted (individually [login] or by site [IP number]); (3a2) unrestricted; (3b) download access (or other in toto access): (3b1) restricted (individually [login] or by site [IP number]); (3b2) unrestricted.

The ideal would be to have fully representative corpora provided with the maximum possible amount of high-level linguistic annotations and rich metadata, which would be available both via sophisticated online user interfaces and for downloading. There is now an urgent need for integration of the (presently) diverse resources and tools in Språkbanken in a way that also takes into account international standardization work in the field of language (technology) resources. Thus,

Språkbanken will be further developed in the following areas, broadly definable as those dealing with infrastructure components (1–5) and user interface/interaction components (6):

1. Standardization of storage and exchange formats;
2. Standardization of annotation, markup and metadata formats;
3. Addition of uniform linguistic annotations to all the corpora of contemporary Swedish;
4. Addition of metadata to existing resources;
5. Definition of a set of processing components and APIs (Application Programming Interfaces) for these components;
6. Development of a set of user interface components for selecting, browsing, searching, annotating, etc., Språkbanken's corpora and lexicons, as well as up- and downloading texts.

Work is well underway in the project on all of these. One aim is to collaborate with other initiatives whenever feasible; thus, the corpus browser frontend Glossa developed by Tekstlaboratoriet, University of Oslo, is now being adapted for use in Språkbanken. This work will be conducted jointly with Tekstlaboratoriet.

The CLARIN preparatory phase work is seen as so important by an institution such as Språkbanken – whose day-to-day activities will be profoundly influenced by the standards, recommendations, best practices, etc., which emerge from CLARIN preparatory phase work – that Språkbanken has decided to use part of the funding for this national project to participate in the preparatory phase of CLARIN; at the present time, this is one of the best ways of safeguarding the future of Språkbanken.

2.3 Spontal: Multimodal database of spontaneous speech in dialog

This section describes the ongoing Swedish speech database project, *Spontal: Multimodal database of spontaneous speech in dialog*. The project takes as its point of departure the fact that both vocal signals and gesture involving the face and body are important in everyday, face-to-face communicative interaction. Our understanding of vocal and visual cues and interactions in spontaneous speech is growing, but there is a great need for data with which we can make more precise measurements. Currently we have very little

data with which we can measure with precision such important aspects of human communication as the timing relationships between vocal signals and facial and body gestures, or how these gestures vary in spontaneous speech or in different speaking styles.

The goal of the Spontal project is the creation of a Swedish multimodal spontaneous speech database rich enough to capture important variations among speakers and speaking styles to meet the demands of current talk-in-interaction research. An important contemporary trend is the study of everyday spoken language in dialog which has many characteristics differing from written language or scripted speech. Detailed analysis of spontaneous speech can also be fruitful for phonetic studies of prosody and also reduced and hypoarticulated speech. The Spontal database will make it possible to test hypotheses on the visual and verbal features employed in communicative behavior covering a variety of functions. To increase our understanding of traditional prosodic functions such as prominence lending and grouping and phrasing, the database will enable researchers to study visual and acoustic interaction over several subjects and dialog partners. Moreover, dialog functions such as the signaling of turn-taking, feedback, attitudes and emotion can be studied from a multimodal, dialog perspective. In addition to basic research, one important application area of the database is to gain knowledge to use in creating an animated talking agent (talking head) capable of displaying realistic communicative behavior with the long-term aim of using such an agent in conversational spoken language systems. The database will be freely available for research purposes.

60 hours of dialog consisting of 120 half-hour sessions will be recorded. Each session consists of three consecutive 10 minute blocks. Subjects are told that they are allowed to talk about absolutely anything they want at any point in the session, including meta-comments on the recording environment and suchlike, with the intention to relieve subjects from feeling forced to behave in any particular manner. Subjects are informed about the time after each 10 minute block. After 20 minutes, they are asked to open a wooden box which contains objects whose identity or function is not immediately obvious. The subjects may then hold, examine and discuss the objects taken from the

box, but they may also chose to continue whatever discussion they were engaged in or talk about something entirely different. The subjects are all native speakers of Swedish and balanced as to gender and whether the dialogue partners know each other or not. This balance will result in 15 dialogs of each configuration: 15x4x2 for a total of 120 dialogs. Currently (February, 2009), about 25% of the database has been recorded.

In the base configuration, the recordings are comprised of high-quality audio and high-definition video, with about 5% of the recordings also making use of a motion capture system using infra-red cameras and reflective markers for recording facial gestures in 3D. In addition, the motion capture system is used on virtually all recordings to capture body and head gestures, although resources to treat and annotate this data have yet to be allocated.

2.4 SweDia 2000 – A Swedish dialect database

The SweDia database consists of recorded speech from 107 dialects representing the dialectal variation in Sweden and Swedish-speaking parts of Finland. The recordings were made in 1999 by a previous research project, SweDia 2000. Each dialect is represented by twelve speakers representing two generations with an equal number of male and female speakers. Research questions that may be addressed using the data are: What are the laws that govern language development and change? To what extent does internal structural coherence govern the development of dialects? The database has until now primarily been used by the SweDia group and a circle of researchers who have obtained personal copies on hard disks. The goal of the present work is to make the database available to a much wider circle by placing it on an internet server together with other language databases accessible via a common web-based interface. It should be possible to perform searches at syllable-, word- or word sequence levels. A first version of (nearly) the entire database already exists hosted on an IMDI-server at the Centre for Language and Literature at Lund University. The result of a successful search can, for example, be a sound file with the desired items and a time-aligned transcription. It should be possible to listen to it directly or download a file for further analysis. In its present form, only parts of the database

material are transcribed.

A part of the database that comprises informal interviews and semi spontaneous monologues will be simultaneously hosted on a server at Tekstlaboratoriet at the University of Oslo. This part of the database will be combined with data collected by the Scandinavian Dialect Syntax project.

To make the databases fully searchable they will have to be transcribed at the word level. This work is in progress and substantial parts of the material are already transcribed. Simple analysis tools will also be available. To the extent that it is possible they will be designed to run on-line. Additional tools will be offered for download.

2.5 Litteraturbanken

The project described in this section – *Litteraturbanken* (the Swedish Literature Bank; <<http://litteraturbanken.se>>) is different from the others described above, in that it has permanent funding by an independent private funding body, the Swedish Academy.

Litteraturbanken is a public digital repository of classical Swedish literary works in scientifically validated editions. It is slated to grow by approximately 100 novel-length works annually. The relevance to CLARIN of this endeavor is found in the following two circumstances:

1. The technical infrastructure of Litteraturbanken was developed by Språkbanken, which is also responsible for developing this infrastructure and maintaining the Litteraturbanken website in its servers. This means that the work on the technical solutions in Litteraturbanken is part of the work in the project described above in section 2.2;
2. Litteraturbanken is developed with the aim that it can serve as a primary data source for research in a number of disciplines in the humanities and social sciences (e.g., literature, various historical disciplines and sociology), using language technology tools, e.g., in the form of text mining.

3 Conclusion

Even though the Swedish Research Council has not set aside funds explicitly intended for CLARIN work, the projects described in the preceding section together represent a funding of 10.6 million SEK (about 1 million Euro), plus about 2.5 million SEK annually to Litteraturbanken. The re-

sources being realized with this funding will be extremely valuable when CLARIN enters its permanent phase.

Acknowledgments

We gratefully acknowledge the following sources of funding for the work described or mentioned above.

The work in the CLARIN preparatory phase by the Centre for Speech Technology, KTH, and Centre for Languages and Literature, Lund University, supported by CLARIN.

The planning project *An infrastructure for Swedish language technology 2007–2008* (a national collaboration, coordinated by Språkbanken, University of Gothenburg), by the Swedish Research Council's Committee for Research Infrastructures (VR dnr 2006-6763).

The project *Safeguarding the future of Språkbanken 2008–2010* (Språkbanken, University of Gothenburg), supported by the Database Infrastructure Committee of the Swedish Research Council's Committee for Research Infrastructures (VR dnr 2007-7430).

The project *Spontal: Multimodal database of spontaneous speech in dialog 2007–2009* (Centre for Speech Technology, KTH, supported by the Database Infrastructure Committee of the Swedish Research Council's Committee for Research Infrastructures (VR dnr 2006-7482).

The project *SweDia 2000 – A Swedish dialect database 2008–2010* (Phonetics, University of Gothenburg), supported by the Database Infrastructure Committee of the Swedish Research Council's Committee for Research Infrastructures (VR dnr 2007-7432).

Litteraturbanken, supported on a permanent basis by the Swedish Academy.

CLARIN in Denmark – European and Nordic perspectives

Hanne Fersøe

University of Copenhagen
Centre for Language Technology
Copenhagen, Denmark
hannef@hum.ku.dk

Bente Maegaard

University of Copenhagen
Centre for Language Technology
Copenhagen, Denmark
bmaegaard@hum.ku.dk

Abstract

This paper gives an overview of the Danish CLARIN project (funding background, national strategic goals, formation of consortium etc.) including the very important priority of aiming the results of the project at researchers from the wide range of all fields of humanities research which is based on language sources, i.e. not exclusively at researchers in the fields of linguistics and language technology, but with a much broader scope. Secondly, it discusses future perspectives of European and Nordic cooperation.

1 The European context

The European Strategy Forum on Research Infrastructures (ESFRI) initiated its Roadmap Process in 2001, and in 2006 it published the first European Roadmap for Research Infrastructures (RI), which was updated in 2008¹. The Roadmap gave its recommendation to 6 SSH-projects (Social Sciences and Humanities), and the European CLARIN project is one of those 6 projects.

At the European Commission level a funding model for European Research Infrastructure (RI) projects was developed in the 7th Framework Programme, a call was opened for those recommended by ESFRI, and 34 projects are currently running, including 5 SSH projects. In parallel, the European Commission has work in progress on a Council Regulation to provide a legal form for the long-term organization to run the pan-European RIs in the construction and deployment phases.

The participation in the construction and deployment of pan-European RIs must be funded nationally, so the 27 EU member states have agreed to develop national Roadmaps. Currently approximately half of these are available.

2 The National Danish Context

2.1 Funding of Danish RI projects or Danish participation in European RI projects

In parallel with the European interest in research infrastructures, the Danish Ministry of Science, Technology and Innovation commissioned the Danish Council for Strategic research to survey the needs and propose a strategy for future research infrastructures. Their report was published in December 2005.

Following these preparatory strategy papers a call for proposals of RIs was published in September 2007 with a pool of 200 million DKK (€27 million) per year for a period of three years.

2.2 Danish Roadmap

Additionally, the Danish Agency for Science, Technology and Innovation is preparing the national Danish roadmap for RIs in agreement with the ESFRI and the Commission process.

3 The Danish CLARIN project

The Danish CLARIN consortium applied for the equivalent of four million euros and was awarded two million for the three year period 2008-2010 for the construction of a national research infrastructure for the humanities, focusing on material expressed in language (written or spoken) and tools to treat this material. This means that Denmark is not in a preparatory phase parallel to the EU-CLARIN project, but that we

¹ <http://cordis.europa.eu/esfri/>

are actually implementing a national research infrastructure.

3.1 The Consortium

The Danish CLARIN consortium has four universities and four cultural institutions as their members with the University of Copenhagen coordinating the consortium. The members are:

- University of Copenhagen
- University of Southern Denmark
- University of Aarhus
- Copenhagen Business School
- Society for Danish Language and Literature
- Danish Language Council
- The Royal Library
- The National Museum of Denmark

A total of 11 research groups are participating with funding, and a 12th group has joined as of January 2009 as an observer.

With these partners the consortium is very strong and to the point, as it has a good combination of the necessary skills and experience: humanities, language technology, language resources, and computer science. The consortium will collaborate with EU-CLARIN where possible, and particularly strive to learn from and adhere to standards as decided at the European level in order to pave the way for Denmark to be an active partner in the construction and exploitation phases of the European project. One of the national tasks for the Danish CLARIN consortium is to propose a strategy for the exploitation at the national level.

3.2 Strategic project goal

The vision is to create a researcher's toolbox by establishing a number of digital Danish text, speech and visual resources and associated tools and to integrate these resources into a web-based environment for research thus creating a much needed support for Danish humanities and enhance its possibilities for European collaboration.

The Danish CLARIN project is eager to follow standards and recommendations developed in the preparatory phase of the European CLARIN project, as far as possible, but as the European project is a preparatory project, the recommendations may not all be available when they are needed for implementation in the Danish

project. The European CLARIN project is assessing existing standards and recommendations in order to be able to determine a set of CLARIN specific recommendations and standards in areas such as technical architecture, meta data, interoperability, IPR and copyright issues etc. However, the Danish CLARIN project needs to proceed, in order to make sure to be able to deliver the results foreseen at the end of 2010.

For this reason it was vitally important for the consortium to design the work packages in such a way as to be able to deliver as a result not only the technical infrastructure but also as many types of content as possible. This means that the project plan contains activities both to deliver already existing resources and to produce new resources. The project is organized into thematically defined main work packages, namely written language resources, spoken language resources and collections of constructed data. Each main work package is subdivided into a number of sub work packages, and in each of these the participants are in the process of collecting, annotating and otherwise producing and including different types of resources.

3.3 Written language resources

In the main work package *written language resources* six different written language resources will be created and made available through the Danish CLARIN infrastructure.

The Danish CLARIN partner Society for Danish Language and Literature (DSL)²: is responsible for collecting a contemporary general language corpus of 15 million words of annotated Danish text per year (i.e. a total of 45 million words), mainly from newspapers and periodicals. This new corpus will cover the period around 2010, and as such it will be supplementing the existing KorpusDK³ which contains around 56 million words from the periods around 1990 and around 2000, respectively. The corpus annotations will be expressed according to TEI P5 specifications. Apart from KorpusDK, DSL has many other interesting and relevant digital resources, as can be seen on their web page, and as a part of the project some of these will also be made available through CLARIN.

University of Copenhagen, Centre for Language Technology (CST)⁴, together with the

² <http://dsl.dk/>

³ <http://ordnet.dk/korpusdk>

⁴ <http://english.cst.ku.dk/>

Danish Language Council (DSN)⁵ is responsible for collecting an 11 million words corpus of annotated sublanguage texts from the period 2000-2010 from broadly selected domains such as health care and medicine, IT, agriculture, construction, meteorology. The corpus will be based on texts originating from experts and semi-experts and with a targeted readership of semi-experts and laymen. At present no such corpus exists for Danish so the sublanguage corpus will represent a truly new type of resource for scientists to work with, and as such it will constitute a valuable supplement to the general language corpus. To learn more about the general language corpus and the sublanguage corpus, see Halskov (to appear).

Another corpus of sublanguage texts will be collected by researchers from the DUDS⁶ group at University of Copenhagen. They will create a corpus of 250,000 words composed of extracts from non-literary texts for everyman's use from the period 1500 to 1750. The texts will be extracted from rare books only obtainable from The Royal Library in Copenhagen, and they will cover subjects such as ethics and moral issues, geography and topography, history, housekeeping and cooking, medical science, mathematics and astrology, natural sciences, pedagogics, etc. (Fersøe 2008b). The texts will be scanned and OCR recognized and marked up according to the Multi Level Text (MLT) annotation (Ruus 2002) which handles orthographical variation, and which will be the key to searching the corpus.

The domains covered in the Everyman corpus mentioned above could be richly illustrated by the images found in existing collections belonging to the section Danmarks Nyere Tid (DNT)⁷ of The National Museum of Denmark. A group from this unit is responsible for creating a pilot corpus of 8,000 images with associated textual descriptions and for making them available on the platform. After deciding the best way of capturing and annotating all the available information from the associated texts, including which language technologies to use for this, they will select more images. Currently there are 50,000 digitized images to choose from. It is not the task of this project to link the Everyman corpus and the DNT images, but this is a future research project. Furthermore the linking could also be ex-

tended to the Danish Dictionary of Insular Dialects, DID⁸, see further down.

Older literary texts will be represented through the work of the Danish writer and Nobel Prize winner, Johannes V. Jensen. Of his work 50 books will be digitized, OCR recognized and annotated, the latter a task which implies adapting the tools, e.g. the PoS-tagger, to older Danish. DSL is responsible for this work together with the Johannes V. Jensen Centre of the University of Århus⁹. In addition DSL will also be specifying a prototypical lexicon of orthographical variation.

Finally a parallel multilingual resource of at least 20 million words will be collected from available bilingual texts. The work will build on experience gained from previous work carried out by research groups at the University of Copenhagen (Maegaard, Offersgaard et al. 2006). While this previous work focused on older texts, namely *The Snowman* by the famous Danish fairy tale writer Hans Christian Andersen, the new parallel corpus will focus on contemporary texts. The texts will be collected and subsequently aligned and annotated, and focus will be on Danish-English and Danish-German. CST is responsible for collecting, aligning, and otherwise annotating the multilingual corpus and for making it available.

One of the challenges in connection with collecting and making available current written text resources is the copyright issue. The consortium is asking permission from writers, publishers and other categories of text owners, and only texts for which permission can be obtained will be included.

3.4 Spoken language resources

In the main work package *spoken language resources* three different spoken language corpora, one of them including video recordings, will be collected, annotated and made available with a number of associated tools.

A group of researchers from the University of Southern Denmark. USD¹⁰, in Kolding will collect video and sound recordings of 20 hours of naturally occurring interaction, mostly from face

⁵ <http://www.dsn.dk/>

⁶ <http://duds.nordisk.ku.dk/>

⁷ <http://www.nationalmuseet.dk/sw6796.asp>

⁸

<http://dialektforskning.ku.dk/publikationer/oemaalsordbogen/>

⁹ <http://www.nordisk.au.dk/jensen/index>

¹⁰

http://www.sdu.dk/Om_SDU/Institutter_centre/Isk/Centre/SoPraCon.aspx

to face situations. The corpus will be annotated according to the Conversation Analysis methods (MacWhinney and Wagner, to appear) to encode overlap, pausing, prosody, and a wide variety of non-lexical features. In addition to this, parts of the corpus will also be annotated with multimodality coding according to the MUMIN system (Jokinen, and Navarretta et al., 2008) for facial and manual gestures, gaze, posture, and proximity. The corpus will be accompanied by a search engine which allows the data to be searched for interactional features, mainly combinations of verbal material, timing plus features marked in the transcription.

Another spoken corpus will be collected by the researchers from the Danish National Research Foundation Centre for Language Change in Real Time, LANCHART¹¹, at the University of Copenhagen. This group is working with corpora collected over a long period of time, and they are re-interviewing some of the informants that were interviewed earlier in order to be able to compare their language between then and now and thus study language change (Gregersen, 2007). There are, however, various confidentiality restrictions which are making it very difficult – if not impossible – to offer free availability to these corpora, so in the CLARIN context a new small corpus of spoken young Copenhagen Danish will be collected and annotated according to the LANCHART standards. The group will also deliver a tool that can be used for analysis by all researchers who want to handle and study spoken language materials.

The third spoken corpus to be delivered through the Danish CLARIN infrastructure is created at Copenhagen Business School, CBS¹². The corpus text is the Danish PAROLE corpus¹³ of which currently 100,000 tokens exist as sound files in lab quality (Henrichsen, 2007). This corpus will be made available with the sound files and with annotations for PoS, syntactic structures, acoustic measurements, phonetic transcription, and more. These data are unique in Denmark for phonetic studies and speech technology. The data will be extended, revised and re-organized to be made available through CLARIN, and so will the accompanying tools for word-level alignment, verification of phonetic transcription, and acoustically based prosodic analysis.

¹¹ <http://lanchart.hum.ku.dk/>

¹² <http://isvcbs.dk/~pjuel/index2.html>

¹³ <http://korpus.dsl.dk/e-resurser/parole-korpus.html>

3.5 Collections of constructed data

The term ‘collections of constructed data’, or technological resources as they are also called, is a loose definition we have used in the Danish CLARIN project to cover resources that are not collected and annotated as they are, such as e.g. written or spoken corpora, but which are carefully selected data put together as a collection according to a specific set of requirements, such as e.g. dictionaries. In the main work package *collections of constructed data* three different sets of constructed data will be made available.

The Danish WordNet, DanNet¹⁴ (Pedersen, Nimb et al. 2008), will be extended from 35,000 to 70,000 synsets in close collaboration between CST and DSL and according to a set of specifications for inclusion of new vocabulary. The extension, more precisely, consists of generation of the new synsets, placing them in the ontological structure of DanNet, determining DanNet equivalents for Base Concepts from Princeton WordNet¹⁵, and establishing the links to Princeton WordNet. The existing coding tool will be slightly enhanced, and an xml-format will be developed.

Researchers from the Jens Peter Skautrup Centre¹⁶ at the University of Århus have developed Jysk Ordbog¹⁷, which is a rich resource of dialects of Jutland. In the CLARIN project the research group will evaluate the current data base format of the dictionary and subsequently redesign it to fit more appropriately with CLARIN standards and formats before making it available through the infrastructure.

Bringing together different types of dictionary resources is scientifically interesting and has obvious benefits for teaching. In the CLARIN project researchers from CST will bring together DanNet and the Danish computational dictionary, STO¹⁸, and thus highly improve the potential of both as a computerized representation of Danish vocabulary, providing not only lexical semantic information, but also syntax and morphology. The work will be based on the positive results of a pilot project (Pedersen, Braasch et al. 2008), and will comprise about 9,000 words.

The research group from Danish Dictionary of Insular Dialects (DID) mentioned earlier is not a CLARIN partner with funding from the grant.

¹⁴ <http://www.wordnet.dk/>

¹⁵ <http://wordnet.princeton.edu/>

¹⁶ <http://www.jysk.au.dk/index.jsp>

¹⁷ <http://www.jysk.au.dk/jyskordbog/jyskordbog>

¹⁸ http://english.cst.ku.dk/sto_ordbase/

The group, however, is currently working with some technical issues similar to those of Jysk Ordbog, i.e. formats, meta data, data structure and tools, and therefore the Danish CLARIN consortium has invited the DID group to become observers in the work package regarding the constructed data.

3.6 Technical platform

The technical infrastructure of the Danish CLARIN platform is in the process of being specified, and it is still too early to give a more detailed account of these matters. Currently the infrastructure is seen as a digital repository with a web user interface managing:

- Access rights given to users based on user verification mechanisms
- Access rights for users to specific content based on resource profiling
- Search and retrieval facilities
- A personal work space
- Communication facilities

3.7 The future after 2010

One of the management tasks of the Danish consortium is to propose a plan for future operation and exploitation of the Danish CLARIN infrastructure. Key elements for which future funding must be found are on the one hand the technical inclusion of Danish CLARIN into EU-CLARIN, and on the other hand the continued inclusion of new resources on to the national infrastructure. Another challenge will be the dissemination of the usefulness of the infrastructure for a wide range of humanities research areas.

4 European and Nordic Perspectives

The history of language technology collaboration among the Nordic countries goes back to the early days of computational linguistics. The first Nordic summer school in computational linguistics was held in Marstrand, Sweden, in 1972, followed up by Bergen 1973 and Copenhagen 1974. These summer schools have been instrumental in the creation of a Nordic computational linguistic community. Later on the Nodalida conferences were started by “Den Nordiske Samarbejdsgruppe for datamaskinel sprogbehandling” with the first conference in Gothenburg 1977, and as the latest step in this direction we have the

creation of NEALT (Northern European Association for Language Technology) in 2007.

The Nordic collaboration has been very important for the building up of the Nordic computational linguistics communities, not least for preparing for European collaboration.

4.1 Content of the Nordic collaboration

Some Nordic countries have languages that are similar and in this case it is highly recommendable to reuse and accommodate tools, standards etc., wherever possible. E.g. the CST lemmatizer for Danish has been trained for Icelandic and is now being used in Iceland. This kind of collaboration will take place only if information about the existence of language technology tools and methods is available. There are several instruments for knowledge sharing and dissemination: the NorDokNet centres (Fersøe, Rögnavaldsson et al. 2005) were supported by the Nordic Council of Ministers, and even if funding has stopped, the collaboration among the centres survives, albeit at a lower level. Similarly the Nodalida conferences are a great help to disseminate knowledge and support Nordic collaboration.

4.2 Merging of Nordic and European perspectives

CLARIN is a European initiative, and this means that CLARIN will provide everything which the Nordic collaboration provides, just at the larger, European, scale: standards and tools are shared with many more languages, and it is possible to collaborate with many more research groups and to be inspired by many more researchers around Europe.

In a successful CLARIN we see the Nordic and the European perspective merging.

Acknowledgements

This project is supported by the Danish Agency for Science, Technology and Innovation, as well as by all partner institutions.

We thank all participants in the Danish consortium for their contribution to the project.

We also thank all the work package leaders for their work package descriptions, which have served as input particularly to section 3 of this document.

References

Hanne Fersøe 2008a. *The Danish CLARIN Project*. CLARIN Newsletter, number 2, July 2008.

- Hanne Fersøe 2008b. *Knowledge for Everyman from the Renaissance to Modern Times*. CLARIN Newsletter, number 4, December 2008.
- Hanne Fersøe, Eiríkur Rögnvaldsson and Koenraad de Smedt 2005. *NorDokNet - Network of Nordic Documentation Centres. Contacts to future Baltic Partners*. Nordisk Sprogteknologi. Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000 - 2004. København 2005, side 13-23.
- Frans Gregersen 2007. *The LANCHART Corpus of Spoken Danish, Report from a corpus in progress*, in J.Toivanen & P.Juel Henrichsen (eds.): *Current Trends in Research on Spoken Language in the Nordic Countries, Volume II*, Oulu University Press, p.130-143, ISBN 978-951-42-8514-1.
- Jakob Halskov (to appear). *Compiling, annotating and publishing corpora in DK-CLARIN, the Danish incarnation of the pan-European initiative for a common resource infrastructure*. To appear in *Corpus Linguistics 2009*, Liverpool.
- Peter Juel Henrichsen 2007. *The Danish PAROLE corpus - a merge of speech and writing*; in J.Toivanen & al (eds) *Current Trends in Research on Spoken Language in the Nordic Countries, vol II*; Oulu Univ. Press 2007, pp.84-93
- K. Jokinen, C. Navarretta , P. Paggio 2008. *Distinguishing the communicative functions of gestures*. In A. Popescu-Belis and R. Stiefelhagen (eds.) *Proceedings of 5th Joint Workshop on Machine Learning and Multimodal Interaction*, Utrecht, September 2008, Springer, 38-49.
- Brian MacWhinney, Johannes Wagner (to appear): *Transcribing, searching and data sharing: The CLAN software*. To appear in *Gesprächsforschung 2009* (ISSN 1617-1837).
- Bente Maegaard, L. Offersgaard, K.F. Joensen. X. Lepetit, C. Navarretta, J. Pedersen, C. Povlsen. 2006. *MULINCO - Korpusplatform til sprog- og oversættelsesstudier*. Tidsskrift for Universiteternes efter- og videreuddannelse, nr. 7 s. 1-15: E-læring i sprogfag, Danmark.
- Bolette S. Pedersen, S. Nimb, L. Trap-Jensen (2008) *DanNet: udvikling og anvendelse af det danske wordnet*. Nordiske Studier i Leksikografi 9, Rapport fra konference om leksikografi i Norden pp. 353-371, Akureyri, Island.
- Bolette S. Pedersen, A. Braasch, L. Henriksen, S. Olsen, C. Povlsen, 2008. *Merging a Syntactic Resource with a WordNet: A Feasibility Study of a Merge between STO and DanNet*. In *Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC'08)*. European Language Resources Association, 2008. 5 s.
- Ruus, Hanne. 2002. *A Corpus-based Electronic Dictionary for (Re)search*, in EURALEX 2002 Proceedings, pages 175-185.

Nordic co-operation in building the language resource infrastructures

Kimmo Koskenniemi

University of Helsinki
Finland

kimmo.koskenniemi@helsinki.fi

Antti Arppe

University of Helsinki
Finland

antti.arppe@helsinki.fi

Abstract

This paper attempts to identify worthwhile goals when building Nordic language resource infrastructures and the relevant parties who should participate their planning and construction. Finally, some actions are suggested which could move us closer to the goals which have been set.

1 Background

We have a long tradition of Nordic co-operation within language technology (Koskenniemi et al. 2007), including a long series of NODALIDA conferences, the Nordic Research Program 2001-2004, NGSLT, and we now have the NEALT organization which hosts special interest groups such as the SigInfra dedicated to research infrastructures for language resources. Similar co-operation has also been practiced in linguistics, e.g. the Nordisk Forskerakademi (NorFA) summer schools and the Scandinavian Conference of Linguistics (SCL).

The *European Common Language Resource and Technology Infrastructure (CLARIN)* infrastructure entered its EC funded preparatory phase 2008-2010 and is creating frameworks according to which the operational CLARIN could be built. All Nordic and Baltic countries are participating CLARIN in various roles.

In Finland, FIN-CLARIN, a consortium of research institutions involved in linguistics and language technology has been formed in 2007 to strive towards CLARIN objectives at a national level. Currently, FIN-CLARIN encompasses the Universities of Helsinki, Joensuu, Jyväskylä, Oulu, and Tampere, the Research Center for the Languages of Finland (KOTUS/FOCIS), and CSC – IT Center for Science, but the consortium remains open to all other Finnish academic or-

ganizations with an involvement in linguistic research or having language resources and technologies available for such research.

As the first step, the FIN-CLARIN consortium members have conducted in 2008 a survey of linguistic research resources and tools that exist within their organizations. In all, 76 distinct collections of resources have been identified in this survey, for which the key descriptive data, identifying the resource, its content, location, and access requirements are available at the FIN-CLARIN website as well as the general *ad hoc* registry maintained by CLARIN, see (www.clarin.eu/view_resources).

As a second step, the FIN-CLARIN consortium has commissioned from CSC – IT Center for Science a White Paper concerning the various possibilities for setting up a Finnish national Authorization and Authentication Infrastructure (AAI) for language resources, as well as a proposal covering the requirements specifications and actual construction plan for implementing such an infrastructure in Finland. Such an AA infrastructure is the technical bedrock which allows for the potential use of a language resource at any of the participating Finnish organizations according to the Single-Sign-On (SSO) principle, i.e. requiring a user's identification only at one's own Finnish home organization. In practice, this now completed development plan realizes the technical framework of the envisioned CLARIN infrastructure within Finland, and is planned to be fully conformant with the pan-European CLARIN AAI, the kernel of which is planned to be operational already in 2009. As the third step, the FIN-CLARIN consortium has commissioned from CSC the actual construction of this AAI in Finland within 2009.

2 Nordic goals

One important goal of Nordic research infrastructures for language resources is obviously to make language and lexical materials accessible

and usable for all those who need them for research, teaching, language planning or similar purposes. The access and use of existing materials should be facilitated, new materials should be created, and measures should be taken in order to secure maximally free availability of the future materials already when the materials will be created.

Just within the Nordic countries, the CLARIN infrastructure should allow for researchers interested in e.g. the overall state of the Swedish language, i.e. Swedish spoken and written both in Sweden and in Finland, to easily access the language resources currently physically located at several institutions, first and foremost *Språkbanken* (The Swedish Language Bank) in Göteborg, Sweden, CSC – IT Center for Science, Finland, the Department of Scandinavian languages and literature at the University of Helsinki, and the Research Center of the Languages of Finland, regardless of what their home organization currently is. Likewise, the CLARIN infrastructure should allow for researchers in e.g. the Department of Fenno-Ugrian Studies at the University of Helsinki to have easy access to the substantial Sámi resources at the University of Tromsø. In addition to such ease of access, the CLARIN infrastructure aims to provide user-friendly interfaces to aggregate such scattered resources as single virtual corpora, and to conduct the most common search and concordancing operations for researchers lacking extensive skills in language technology and programming, which would be necessary to work by themselves directly with the source format of the resources.

The resources for CLARIN or national language resource infrastructures are limited. In order to proceed fast and get the appropriate high quality services available, the Nordic participants now have an opportunity to get more by smart division of labor and by co-ordination, making the most of the current individual strengths of all the parties.

This paper also discusses how the Nordic countries could better integrate themselves in the European CLARIN which is, of course, the best, if not the only way to offer the Nordic researchers the access to materials and tools in other EU countries.

3 Actors

It is important to get the relevant parties involved, including but not restricted to:

- researchers in various disciplines such as linguistics, language technology, or machine learning who need linguistic materials in their research and who sometimes produce new materials,
- researchers in other disciplines who in fact essentially work with linguistic data, e.g. historians, sociologists, or theologians, just to mention a few fields,
- funders of research projects who can require allowing free access, and compliance with standard formats as new materials are produced as a result of the projects,
- specialists in language planning or language cultivation (*språkvård*), who utilize the materials in their work and compile new dictionaries, norms for language users, and compile new corpus materials,
- commercial parties such as publishers and broadcasting companies who own or possess written and spoken materials, as well as language technology companies who need written or spoken corpus materials and create language technology tools using these materials,
- libraries, museums, and some commercial companies such as Google and Microsoft Corporation which may have huge archives of materials and which are involved in digitizing and storing these archives,
- organizations of authors and journalists, as well as the organizations which process the copyright fees of authors and performers, and
- experts in copyright legislation.

There is an obvious need for attracting relevant parties to the work because relevant materials exist and are controlled by them. In addition, risks will increase if those parties are not motivated and co-operative.

At first sight, some of these parties might appear to have conflicting interests. It would be nice for the researchers if they could use all published materials on an open access basis. This might, however, conflict with the legitimate commercial interests of the publisher if they intend to print and sell copies of such a work. We think that there may still be workable compro-

mises where the commercial publisher can feel comfortable and safe at the same time as the researcher can use the texts and other language materials fairly freely. In order to find and establish such practices, one definitely needs contacts, discussions, and negotiations, and in the long run, relatively permanent, organized fora through which such activities take place. Importantly, establishing relations of trust between the various actors requires extensive engagement and time.

4 Organizing Nordic co-operation

Probably the best and only truly operational basis for Nordic co-operation with language resource infrastructures would be based on *national infrastructure consortiums* which are anyway needed in the CLARIN framework. They will be the essential *primary* parties in applying for national funding and in setting priorities for tasks and steps in building resources and the infrastructure.

The European CLARIN will neither build nor fund the national or regional CLARIN centers, and the European CLARIN will not build the materials for national languages. These tasks have to be funded and carried out nationally, and most likely through some national consortium which represents the most relevant parties.

SigInfra of NEALT is a special interest group dedicated for the advancement of Nordic co-operation in language resource infrastructures. SigInfra cannot, however, assume alone much of the responsibilities of building the national infrastructures. But SigInfra, together with national consortia, definitely can make the building of CLARIN compatible resources and centers much more successful.

In a nutshell, the organization could consist of national language resource consortia and a board consisting of one or two representatives nominated by each consortium.

5 Forms of co-operation

Let us suppose that there is a national consortium in each country which is building a national infrastructure for language resources. If so, that would provide an excellent basis for Nordic co-operation aiming at the integration of the national infrastructures into mutually compatible CLARIN nodes. Simply put, a board consisting of representatives from those consortia would plan, co-ordinate, and synchronize the common activities. The national consortia would then carry out the actual tasks which have been agreed upon.

The board could e.g.

- co-ordinate the collecting of certain information by the participating member consortia (such as an inventory of national digital text, speech and lexical materials),
- co-ordinate the application for any national funding and the implementation of the (successful) funding decisions, and store and make the results available as needed,
- initiate discussions and possible negotiations concerning the optimal selection of institutions and centers for various CLARIN service centers, along with the co-operation and division of labor between present or future CLARIN service provider centers,
- discuss and provide recommendations on types and levels of CLARIN metadata describing the language materials,
- discuss and co-ordinate producing, enhancing and sharing of software tools to become parts of CLARIN resources or services,
- apply for Nordic funding for arranging meetings about Nordic language resource infrastructures,

The board would have no resources and practically no funding of its own. All work would be carried out with the funding of national consortia and by their staff. Therefore, the adequately funded national research infrastructure consortia are crucial.

6 Expected results of the co-operation

There are many kinds of small or important results or benefits that could be achieved with Nordic co-operation.

One of the significant and achievable goals would be that the CLARIN infrastructure in the Nordic countries could become operational earlier for the benefit of the research community, than might happen otherwise as a result of national activities altogether non-dependent of each other. A common effort might have a better opportunity of getting adequate local funding. The co-operation might also help national efforts to find better practices and avoid poor design and miscalculations, and learn from the experiences of organizations which have a chance to try out

the construction of some service first. For instance, the forthcoming Finnish experiences in setting up a national AAI for language resources ought to be disseminated to all other Nordic consortia.

One equally important goal would be that the implementation of a good functional Nordic CLARIN might be less expensive to build. This could result from the division of labor where partners concentrate their efforts in components where they have special expertise, and reuse parts which others have created, or simply benefit from the prior experiences of other partners. Moreover, we must note that there is no overabundance of qualified technical people available with the necessary skills to implement the infrastructure. Computing centers at universities and national research institutions may have a critical mass of such people, but these are almost always already quite extensively involved in a range of support activities, providing services to many scientific fields – CLARIN is not the only research infrastructure on the European block. Once we are able to secure on a stable basis such human resources in some organization in a Nordic country, and in addition establish a good working relationship with such an organization to cater to CLARIN needs, we might as well make the most of this capacity throughout all the Nordic countries.

Since the technological environment in which CLARIN operates is dynamic, we must be prepared for changes in the infrastructure as it eventually emerges. For instance, achieving co-operation among the currently existing national authentication (identification) federations requires a relatively extensive network of mutual agreements. However, it is possible that within a few years we will have only one pan-European identity federation to provide authenticated user identities throughout Europe. Nevertheless, in the meantime we have to make do with what exists now. Nevertheless, this does not entail that the first used solutions and the organizations that provide them will be permanently fixed. To the contrary, CLARIN is fundamentally a distributed research infrastructure, allowing and requiring for the moderate duplication of resources and services – and their gradual development and improvement – in order to guarantee maximal operability.

Present archives of digital language materials are somewhat scattered. Whereas the acquisition and license management necessarily involves many institutions, the data processing of lan-

guage materials is mostly modest. Even 10^{12} words of text materials is manageable, and the processing and searching of such masses is not a real problem. On the other hand, managing standardized and high quality data security, state of the art authentication and authorization and metadata harvesting might consume a significant portion of the personal resources at some relevant centers. Nevertheless, the challenge that we must solve in using in some cases a smaller set of centers or even a single one to provide some particular service is that the normal end-users in each Nordic country are entitled to receive an equally high level of support in using and relying on such service, regardless of what their affiliation is.

7 Conclusions

We urge that Nordic organizations with linguistic resources and tools formally establish national CLARIN consortia within each Nordic country. If and when such are already existing, we encourage that they be extended to include all relevant national organizations, and that these organizations be also encouraged to become members of CLARIN at the European level. At the same time, we propose that these Nordic national consortia formally establish a forum/organ for pan-Nordic co-operation and settle on principles guiding this co-operation. It is our firm belief that such co-operation and co-ordination of Nordic CLARIN activities will be of substantial benefit to all involved parties.

References

Kimmo Koskeniemi, Krister Lindén and Torbjørn Nordgård (editors). 2007. *Expert Panel Report: The Nordic Countries, A Leading Region in Language Technology*. Publications, No. 44, Department of General Linguistics, University of Helsinki.

Two decades of Lithuanian HLT

Rūta Marcinkevičienė
Vytautas Magnus University
Kaunas, Lithuania
ruta@hmf.vdu.lt

Abstract

This paper aims at a short overview of the development of the Lithuanian language resources infrastructure in the last two decades in the context of European cooperation. It also presents national policies related to research infrastructures and suggests possible joint activities on different levels, such as European, institutional and personal.

1 Introduction

Baltic languages experienced as many changes during the 20th century as during the whole span of their autonomous existence after separation from their common root, i.e. the proto-Baltic dialect. The biggest challenge for their survival after the appearance of their written and printed variety is their computerisation and utilization in HLT (Marcinkevičienė 2006). The last two decades of the 20th century were important as a number of HLT related activities were performed:

- localisation of general tools,
- digitalisation (including adaptation of digitalised resources),
- compilation of tools, language resources and knowledge bases,
- training and research,
- documentation and publicising.

The first two types of activities, i.e. localisation of the user interface and digitalisation of cultural heritage cannot be classified under HLT proper. However, some types of digitalised products can be used as linguistic resources, e.g.

- Database of Old Lithuanian Writings (<http://www.lki.lt/seniejirastai>),

- Dictionary of Lithuanian Language (<http://www.lkz.lt>),
- Dictionary of Contemporary Lithuanian Language (<http://www.lki.lt/dlkz/>),
- Dictionary of Toponyms (<http://lkz.mch.mii.lt/Vietovardziai>),
- Database of Lithuanian Dialects (<http://tarmes.mch.mii.lt>).

However, digitalised resources are of limited use as resources, therefore a greater prominence is given to the third type of activity, i.e. compilation of general and special corpora and language processing tools.

2 Short overview of Lithuanian HLT

Resource development in Lithuania as in many other countries started with the development of its first corpus. The impetus for that was based on a one-term stay at Stockholm University financed by a scholarship of the Swedish institute in 1991. During that stay knowledge was acquired about the corpus of the Swedish language. The idea of compiling such a corpus for the Lithuanian language was then introduced at the recently reopened Vytautas Magnus University in Kaunas and supported by its administration. As an outcome the Centre of Computational Linguistics (CCL) started in 1994. Before that there were a few personal initiatives in that direction. One of them was the construction of a lemmatiser and a morphological analyzer. Another initiative, the Dictionary of Word Frequencies, was carried out by a group of scholars supported by the Lithuanian State Science and Studies Foundation. The dictionary was based on a one million word corpus which was not exposed to public use.

The CCL as a department was open to a wide range of possibilities to participate in the resource building activities promoted by EU at that time. I would like to mention the most important

moments for the development of the Lithuanian HLT:

a) participation of the CCL in the ECI (European Corpus Initiative) project by way of supplying a modest amount of Lithuanian texts, marked up according to TEI-conformant mark-up language (1993).

b) A long term engagement of the CCL in the project meant to build Trans-European language resource infrastructure, named TELRI (1995-2001). It offered a possibility for an extended collaboration for participants from more than 20 countries, mostly Central and Eastern European, who had never participated in EU projects before. The most useful activities at that time were the co-operation in compiling parallel multilingual corpora, text archives, translating bridge dictionaries, building or adapting software tools, and on the top of it all, acquiring a know-how and theoretical approach to the compilation and exploitation of national language resource infrastructure. TELRI offered a forum for discussions and presentations of resource-based research at its annual seminars as well as at numerous meetings and in newsletters. Besides, it attempted to register all the institutional participants such as language organisations, research institutes, and events (conferences, schools, seminars, etc.) in the field of resource infrastructure of that time. That particular TELRI activity overlapped with and supplemented ELSNET.

c) Last but not least participation of the national program "Lithuanian language in the Information Society 2000-2006" has to be mentioned. The most obvious outcome of the programme for the Lithuanian HLT was compilation of the corpus of 100 million running words and some tools (e.g. corpus query system and collocation extraction tool, a system of morphological annotation and disambiguation) open for public use at <http://donelaitis.vdu.lt>.

Thus, combination of both national and European projects enabled creation of the first tools and resources for Lithuanian. Without EU initiatives national projects and programs would have been hardly possible.

Later developments in the field financed mostly by national foundations ended up in production of the following tools and resources:

- a morphologically annotated corpus (115 million running words),
- an annotated manually checked corpus of one million words,
- a set of parallel corpora:
 - . a bidirectional Czech-Lithuanian and Lithuanian-Czech corpus of five millions words
 - . English-Lithuanian corpus of 18 million words in size,
- a database of Lithuanian nominal collocations, extracted from the corpus of 100 million words,
- a number of tools such as
 - . a tool for the automatic identification of text functions for the Lithuanian language,
 - . the tool for the extraction of collocations,
 - . a Lithuanian tagger,
 - . the Aligner2067,
 - . an automatic accentuation tool for the Lithuanian language,
 - . a corpus of Spoken Lithuanian language,
 - . a universal annotated database of speech recordings.

Above, we confined ourselves to the tools for language resources made at Vytautas Magnus University and sponsored mainly by two national funding agencies, i.e. Lithuanian State Language Commission and the Lithuanian State Science and Studies Foundation.

Other institutions developed a set of tools and databases for public use or purchase. The State Commission of the Lithuanian Language is monitoring an open terminological database <http://terminai.vlkk.lt>. Institute of Mathematics and Informatics digitalised term dictionaries from 27 branches into one database <http://www.terminynas.lt/>. A private company *Fotonija* is known for its electronic dictionaries of international words *Interleksis*, *TŽŽ*; English-Lithuanian dictionaries *Alkonas* and *Anglonas*, French-Lithuanian dictionary *Frankonas* and a spellchecker *Juodos avys* <http://www.fotonija.lt/>. A corpus of academic discourse has been started at Vilnius University, Faculty of Philology.

The most recent jointly developed tool was a rule-based machine translation system for the translation of English internet texts into Lithuanian <http://www.vertimas.vdu.lt>. It was developed by a group of companies among which

Prompt (St. Petersburg), Fotonija (Vilnius), Ala Software (Kaunas). They co-operated within the framework of a project financed by EU Structural Funds. At the moment this machine translation tool is the most popular tool for the Lithuanian language and it is used for the translation of circa 2 millions texts per month by 40,000 registered and 600,000 occasional users. Before the automatic machine translation system there was an automatised translation tool *Vertimo Vedlys* incorporated in text editor *Tildės biurais* <http://www.tilde.lt/> together with a spellchecker and multilanguage support software. It translates NPs and simple sentences.

According to Sarasola's typology of language technology resources (Sarasola, 2000), the Lithuanian language resources, as they are at the moment, consist of

- a) so-called foundations, i.e. raw corpora, machine-readable dictionaries, speech databases,
- b) basic tools such as statistical tools for corpus treatment, a morphological analyzer, generator and lemmatizer, and a speech recognition system dealing with isolated words,
- c) medium-complexity tools such as spell checkers and a structured lexical database which includes multiword lexical units.

Advanced tools, however, do not exist for Lithuanian HLT. Such tools include

- syntactically annotated corpora (treebanks),
- grammar and style checkers,
- lexical-semantic knowledge bases or concept taxonomies such as WordNet,
- word sense disambiguators,
- speech processing tools functioning at sentence level.

On top of those tools there still is, according to the hierarchy of Sarasola, the category of the most sophisticated resources, the so-called multilinguality and general applications. These include:

- semantically annotated corpora,
- information retrieval and extraction,
- dialogue systems,
- language learning systems,
- machine translation.

The latter was recently developed by a co-operation from a group of companies (see above), but the others are not present in Lithuanian HLT.

The question is whether it is possible to adapt the existing advanced tools, made for other languages, and to avoid reinventing a wheel. Our rule-based MT system was immediately followed by the appearance of a stochastic tool presented by Google. If known in advance, compilation of a rule-based MT system could have been postponed as from the point of view of a small language, duplication of tools is a waste of time. However, since the stochastic tool is of a worse quality, it is worthwhile to have a rule-based MT system. Moreover, it is desirable to develop it into a bidirectional translation system and add the Lithuanian-English component. In general, we are of the opinion that compilation of language specific tools is to be strived for based on universal tools and adapt them to our language. However, in cases where so-called universal and language independent tools are based on the prevailing language probabilistic models (usually for English) such tools are mostly not usable for easy generalization towards other languages (cf. Borin, 2004).

3 National policies related to research infrastructures

On a national level research and development programs continue to promote HLT related activities. The Ministry of Education and Research is responsible for the second phase of the program *Lithuanian Language in the Information Society 2010-2015* that deals with localisation, resource and tool creation, documentation and some other activities. The Lithuanian Research Council has launched the first national program *Heritage and Identity* that encompasses digitalization of intangible heritage. Recently, language digitalization is also stimulated in a wider program on specific Lithuanian cultural and philological trends *Lituanistikos plėtra 2009-2015*.

The most important development and support of resources is foreseen in the framework of the National Research Infrastructure (NRI) compatible with ESFRI requirements for national states. The strategy of NRI includes documentation and unification of existing national resources as well as support for trans-national initiatives such as CLARIN, CESSDA and other similar joint infra-

structures for the Social Sciences and Humanities (SSH). National support for research infrastructures in general and HLT in particular is timely since "SSH researchers rely on new technologies, and real overhead costs for SSH research have increased dramatically over the past 20 years, without government subsidies necessarily reflecting these changes. Consequently, more and more SSH research depends on capital injections to develop cutting edge data sets and develop retrieval systems" (METRIS report 2009).

It can be concluded that most of Lithuanian HLT related activities, mentioned in the Introduction, are taken care of on national level. Training and research, however, remain the least attended activities. Fundamental or applied research on computational and corpus linguistics, artificial intelligence and a number of fields can be carried out within the scope of national and EU programmes. Training is in the worst position with one BA and one MA level programs both in the Faculties of Humanities at Kaunas University of Technology and Vytautas Magnus University respectively. The lack of post-graduate studies in fields related to HLT was partially covered by the courses and other activities offered by the Nordic Graduate School of Language Technologies, one of the most fruitful initiatives in the history of Baltic and Nordic co-operation in the field.

4 General considerations

The experience of building a national language resource infrastructure gained while participating in various enterprises during almost 20 years gives some basis to evaluate existing forms of co-operation on:

- EU level,
- transnational,
- research communities,
- national,
- institutional,
- personal.

The most fruitful seem to be the forms of long-term institutional participation in EU or transnational bodies that are supported and sponsored by the state. Therefore, such bodies as CLARIN are most promising in the long run. However, the scope of the enterprise is so big that it may prevent its participants from their involvement in

smaller groups and communities. Thus the idea of Nordic-Baltic unit in the framework of CLARIN is mostly welcome, especially if it is supported by national research funding agencies pooling their effort on both policy making and specifically supporting levels.

Lithuania would be interested in exchange of its resources into adaptable tools or in participation in large scale pan-European infrastructural projects. Joint documentation efforts, training of researchers aiming at joint degrees from co-operating universities, and common research infrastructures are a few possibilities to be mentioned. In general, official or institutional levels of co-operation is a precondition for further development carried out mostly on personal and research community level. The latter, either national or international, is the best medium for spreading ideas, offering new tools and methods of research for colleagues from different fields. A good example of such co-operation could be the compilation of corpus-based ontology of computer security and dependability terms (ulo et al., 2007). The HLT community is one of the numerous groups, therefore it would be of paramount importance to engage other formal or informal SSH groups around the Baltic Sea that deal with linguistic resources. That can be carried out via personal overlapping participation in CLARIN and international associations, e.g. International Pragmatics Association or Societas Linguistica Europaea to mention just a few. Therefore further networking is a field of obvious European added-value.

References

- Oliver ulo, Gintar Grigonyt , Merylyne Hernandez, Algirdas Avizienis, Johann Haller, R ta Marcinkevi ien . 2008. Building a Thesaurus of Dependability and Security: a Corpus Based Approach. *Proceedings of the Third Baltic Conference on Human Language Technologies*, October 4-5, 2007, Kaunas, Lithuania, 71-78.
- Lars Borin. 2004. Language technology resources for less prevalent languages: will the Münchhausen model work? *Nordisk Sprogteknologi*, 2003. København, Denmark, 71-82.
- METRIS. 2009. *Emerging Trends in Socio-economic Sciences and Humanities in Europe*. The METRIS Report.
- R ta Marcinkevi ien . 2006. Balt kalb išlikimo problema informacin je visuomen je (The problem

of survival of Baltic Languages in the information society), *Prace Baltystyczne* 3. Warsaw, Poland, 37-43.

K. Sarasola. 2000. Strategic priorities for the development of language technology in minority languages, LREC 2000, *Proceedings of the Second International Conference on Language Resources and Evaluation "Developing language resources for minority languages: reusability and strategic priorities"*, Athens, Greece. ELRA. 106-109.

Estonian language technology Anno 2009

Einar Meister

Institute of Cybernetics at
Tallinn University of
Technology
Tallinn, Estonia
einar@ioc.ee

Tiit Roosmaa

Department of Computer
Science
University of Tartu
Tartu, Estonia
tiit.roosmaa@ut.ee

Jaak Vilo

Department of Computer
Science
University of Tartu
Tartu, Estonia
jaak.vilo@ut.ee

Abstract

The paper will give an overview of developments in Estonia in the field of Human Language Technologies. Despite of the fact that Estonian is one of the smallest official languages in EU and therefore in less favourable position in the HLT-market, the national initiatives are undertaken in order to promote HLT development in Estonia.

1 Introduction

The development efforts of human-computer interaction during the past few decades have been directed towards natural communication using spoken language input and output. For several, especially "big" languages, progress in language technology has been impressive - research results have been successfully exploited in commercial products and services, and the HLT-market shows growing trends. According to the Euromap report (Joscelyne, Lockwood, 2003) on HLT progress in EU countries, the leading positions are held by the UK, Germany, France, the Netherlands and Finland. In the case of the first three countries it can be explained mainly by large market demands, whereas in the latter cases the leading position has been achieved due to several simultaneous factors - healthy environment for R&D, relatively large and strong research community and significant national-level support in the HLT area. Although linguistic and cultural diversity are the core values of the EU and discrimination based on language is prohibited by the EU's charter of fundamental rights (article 22) we need to face the fact that there are primary, secondary and even tertiary languages of

commercial relevance (TC-STAR report, 2006). Development of HLT tools for a new language is a more or less fixed effort and does not correlate with the number of speakers; therefore the smaller languages are in less favourite position, as the costs per capita for HLT development will be higher. What should be done for smaller languages in order to strengthen their market positions and survival in a multilingual EU? - these are crucial questions for smaller countries and also for EU language policy makers wanting to prevent Gutenberg's effect from taking place in the computer age. These issues have been addressed in Krauwer's papers (2005, 2006). Krauwer's claim that the strong industrial bias of EU programmes has led to the situation where the major part of HLT funding is used to support a few major EU languages seems to hold true. As there are not many options (due to the subsidiarity principle) to get financial support from the EU for the technological development of smaller languages, activities on the national level are of great importance. In Estonia several activities to promote R&D in HLT area have been undertaken during the last decade. Mostly these activities have been initiated by the academic groups working on HLT-related topics; in parallel with academic research a lot of effort has been put into explaining the role of HLT in the information society. Although not all initiatives were fully successful, they played an enlightening role among decision-makers and contributed to the forming of a positive attitude in the society. As a result of the joint effort of researchers and the Ministry of Science and Education, the National Programme for Estonian Language Technology (2006-10) was launched. In this paper we will share our experiences in promoting HLT-related

national activities and introduce the Estonian HLT roadmap as well as on-going R&D projects.

2 HLT research in Estonia

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

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

(1) **University of Tartu**, represented mainly by the **Research Group on Computer Linguistics**

(<http://www.cl.ut.ee>). Their research areas cover:

- formal descriptions of morphology, syntax and semantics of the Estonian;
- creating Estonian language resources: electronic corpora of written and spoken language, dialogue corpora, parallel corpora, lexical and semantic database (thesaurus, Estonian WordNet);- software development for morphological, syntactic and semantic analysis and synthesis.

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

(2) **Institute of the Estonian Language, Research Group on Language**

Technology (<http://www.eki.ee>), focused on:

- rule-based morphological systems: formal grammars and software (morphological synthesis and analysis, morphological disambiguation);
- language resources: electronic versions of traditional dictionaries, linguistic databases, text-based dictionaries, lexicons for machine translation, www-applications;
- phonetics and speech technology:

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

(3) **Institute of Cybernetics at Tallinn**

University of Technology represented by the **Laboratory of Phonetics and Speech Technology**

(<http://www.phon.ioc.ee>). Its R&D activities include:

- experimental phonetics: research on Estonian sound system and prosody including Estonian as L2;
- speech technology: speech analysis and speech synthesis, automatic speech recognition (ASR);
- speech databases: Estonian BABEL, Estonian SpeechDat, etc.

There also exist a few small private HLT companies:

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

Keelevara

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

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

TEA Publishers (<http://www.tea.ee>) - established in 1991, one of the leading publishers of economics dictionaries and

foreign language textbooks in Estonia. **Imprimaatur** - founded in 1996, offers consulting, training and quality assurance services related to translation and term banks.

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

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

2.1 HLT financing

Reforms of research funding in the beginning of the 1990s mark a new era for the academic community in Estonia. A competition-based funding scheme was introduced where all research fields had to compete for survival. HLT research groups survived quite well due to successful participation in several international projects (e.g. EU Copernicus). Starting at the end of the 1990s, additional funding sources were opened: the Estonian Language Technology programme initiated by the Estonian Informatics Centre (1998-2000). Within this programme the first Development Plan for Estonian Language Technology was compiled in 1999;

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

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

3 Estonian HLT Roadmap

The roadmap (Figure 1) compiled in 2004

shows the baseline - the resources and tools developed in Estonia during several years before 2004, and presents the future developments in three major action lines:

Action Line 1: Spoken Language Technology including:

- speech synthesis: creating Estonian TTS software and development of an audio-visual synthesis prototype;
- speech recognition: creating a prototype of limited vocabulary ASR and development of language-specific methods for unlimited vocabulary ASR;
- dialogue systems: creating limited-domain intelligent services capable of replacing routine human work.

Action Line 2: Written Language Technology including:

- language processing methods: formalisms for automated processing of different language levels (morphology, syntax, semantics, pragmatics), modeling and creating of corresponding prototypes;
- machine translation: create methods for translating to and from Estonian, compile multilingual vocabularies and mechanisms of transforming syntactic structures; develop prototype for Estonian <-> English machine translation.

Action Line 3: Language Resources including:

- creating infrastructure for collection and management of different language resources;
- collecting different types of resources: speech and text corpora, and electronic dictionaries.

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

4 Towards national HLT programme

In 2003 the Development Strategy of the Estonian Language 2004-2010 was compiled by the members of the Estonian Language Council and was approved by the Estonian Government on August 5, 2004.

(http://www.eki.ee/keelenoukogu/strat_en.pdf)

The strategy provides a research-based description of the situation of the Estonian

language, the objectives that need to be achieved, the necessary steps and institutions and people involved. The development plan of the Estonian language covers all the major areas of language use including language technology.

4.1 National Programme for Estonian Language Technology (NPELT)

NPELT

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

The main goal of NPELT is to develop technology support for the Estonian language to the level that would allow functioning of Estonian in the modern information society. NPELT is funding HLT-related R&D activities including creation of reusable language resources and development of essential linguistic software (up to the working prototypes) as well as bringing the relevant language technology infrastructure up to date. The resources and prototypes funded by the national programme are declared public.

NPELT management is carried out by a steering committee of 9 members (including HLT experts and representatives of the ministries), and a programme coordinator. Responsibilities of the steering committee include the evaluation of project proposals and progress reports, making funding proposals, purposeful use of public funding, surveying the developments in the HLT field on the national and international scale, etc. General rules adopted by the committee:

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

on licence agreements.

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

On-going projects: In 2009, 23 projects have been funded (2006: 17, 2007: 20, 2008: 23) which cover a wide range of topics (see <http://www.keeletehnoloogia.ee/projektid>):

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

5 Centre of Excellence in Computer Science

Estonian language technology researchers are also engaged in the Estonian centre of excellence EXCS (Estonian eXcellence in Computer Science) to be financed over the period 2008-2015. The general objective of the centre of excellence, composed of the research staff of Institute of Cybernetics at the Tallinn University of Technology, Cybernetica AS and the University of Tartu and representing a major part of the computer science research conducted in Estonia, 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 their research results on academia and industry-society as well as to popularize them, 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. According to the Estonian R&D strategy “Knowledge –Based Estonia 2007-2013”, ICT are one of the key technologies for the Estonian RD&I.

6 Centre of Estonian language resources

In 2008, a project of setting up the Centre of Estonian language resources at the University of Tartu was started in the overall framework of the national programme “Estonian Language Technology”.

The natural language resources can be used by different end-users only if the existing resources are well-documented, archived and publicly accessible. In order to support such activities which sometimes may seem gratuitous from the point of view of language resource creators, there need to be a fixed infrastructure to manage and coordinate these activities in Estonia, starting from elaborating the corresponding language technology standards up to drawing the contracts/licence agreements necessary for the use of these language resources.

To achieve this goal, an ESFRI project CLARIN (Common Language Resources and Technology Infrastructure, <http://www.clarin.eu>) has been launched. The University of Tartu is the official representative of Estonia among the 31 partners of CLARIN. The participation in the CLARIN network provides a unique opportunity to involve the pan-European experience in solving our problems.

A similar project titled “Language Technology Documentation Centre” (<http://www.nordoknet.org/>) started in the Nordic countries in 2002 under the auspices of the Nordic Council of Ministries. That Centre has been instrumental in creating a network of centres in Finland, Sweden, Denmark, Iceland and Norway.

The Centre of Estonian language resources will do utmost 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, e.g. linguists, teachers, creators of software systems and their applications, civil servants, etc.

7 Conclusions and future prospects

The national programme has created favourable conditions for HLT development in Estonia. Obviously not all HLT fields are equally addressed and it would be naive to expect that all essential prototypes and resources will be created within a short period. The steering committee is planning an update of the HLT roadmap and takes the initiative towards defining a BLARK (Basic Language Resource Kit) for Estonian.

8 References

- Joscelyne, A., Lockwood, R. (2003). Benchmarking HLT progress in Europe. The EUROMAP Study. Copenhagen 2003.
- Krauwer, S. (2005). How to survive in a multilingual EU? *Proc. of The Second Baltic Conference on HLT*, April 4-5, 2005, Tallinn, Estonia, pp. 61-66.
- Krauwer, S. (2006). Strengthening the smaller languages in Europe. *Proc. Of 5th Slovenian and Ist International Language Technologies Conference*, October 9-10, 2006, Ljubljana, Slovenia. Retrieved on 11/6/2007 from http://nl.ijs.si/is-ltc06/proc/01_Krauwer.pdf
- TC-STAR report (2006). Human language Technologies for Europe. Retrieved on 10/12/2007 from http://www.tc-star.org/publicazioni/D17_HLT_ENG.pdf

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

Figure 1. Estonian HLT Roadmap for 2004-2011

Icelandic Language Resources and Technology: Status and Prospects

Eiríkur Rögnvaldsson University of Iceland eirikur@hi.is	Hrafn Loftsson Reykjavík University hraf@ru.is	Kristín Bjarnadóttir Árni Magnússon Institute for Icelandic Studies kristinb@lexis.hi.is	Sigrún Helgadóttir Árni Magnússon Institute for Icelandic Studies sigruhel@hi.is
Matthew Whelpton University of Iceland whelpton@hi.is	Anna Björk Nikulásdóttir University of Iceland abn@hi.is	Anton Karl Ingason University of Iceland antoni@hi.is	

Abstract

We give an overview of Icelandic language technology since its inception ten years ago and describe briefly its main achievements. Then we outline the research program of the Icelandic Language Technology community for the next few years, which is being implemented thanks to a large grant which has just been allotted to the program by the Icelandic Research Fund. Finally, we discuss the need for Nordic cooperation within Language Technology and put forward some concrete proposals for enhanced cooperation.

1 Introduction

Ten years ago, Icelandic language technology (henceforth LT) was virtually non-existent. There was a relatively good spell checker, a not-so-good speech synthesizer, and that was all. There were no programs or even individual courses on language technology or computational linguistics at any Icelandic university, there was no ongoing research in these areas, and no Icelandic software companies were working on language technology.

All of this has now changed and Icelandic language technology has been firmly established. In the fall of 1998, the Minister of Education, Science and Culture appointed a special committee to investigate the situation in language technology in Iceland and come up with proposals for strengthening the status of Icelandic language technology. The committee handed its report to the Minister in April 1999 (Ólafsson et al., 1999) and in 2000, the Government launched a special Language Technology Program (Arnalds, 2004;

Ólafsson, 2004), with the aim of supporting institutions and companies in creating basic resources for Icelandic language technology work. This initiative resulted in several projects which have had profound influence on the field (cf. Rögnvaldsson, 2008).

In this paper, we will first give an overview of this work and other activities in the field during the past ten years. Then we will briefly outline the research program of the Icelandic LT research community for the next few years and point out the importance of open source policy for less-resourced languages. Finally, we will discuss the importance of Nordic cooperation within LT and put forward some concrete proposals to this effect, especially concerning education and dissemination of information.

2 Icelandic LT Work 1999-2009

In the report of the Language Technology Committee (Ólafsson et al., 1999), four types of actions were proposed in order to establish Icelandic language technology:

- The development of common linguistic resources that can be used by companies as sources of raw material for their products.
- Investment in applied research in the field of language technology.
- Financial support for companies for the development of language technology products.
- Development and upgrading of education and training in language technology and linguistics.

This has all been done, to some extent at least (Rögnvaldsson, 2008). The main direct products of the LT Program are the following:

- A full-form morphological database of Modern Icelandic inflections (Bjarnadóttir, 2004, 2005).
- A balanced morphosyntactically tagged corpus of 25 million words (Helgadóttir, 2004).
- A training model for data-driven POS taggers (Helgadóttir, 2005, 2007).
- A text-to-speech system (Rögnvaldsson, Kristinsson and Þorsteinsson, 2006).
- A speech recognizer (Rögnvaldsson, 2004; Waage, 2004).
- An improved spell checker (Skúlason, 2004).

After the government-funded LT Program ended, researchers from three research institutes (University of Iceland, Reykjavik University, and the Árni Magnússon Institute for Icelandic Studies) decided to join forces in a consortium called the *Icelandic Centre for Language Technology* (ICLT), in order to follow up on the tasks of the Program. The ICLT serves its role by:

- maintaining an information center for Icelandic language technology by running a website (cf. Rögnvaldsson, 2005);
- encouraging cooperation on LT projects between universities, institutions and private companies;
- organizing and coordinating university education in language technology;
- taking part in Nordic, European and international cooperation in the field of language technology;
- initiating and participating in research projects in language technology;
- initiating and participating in commercial projects in language technology;
- keeping track on resources and products in the field of language technology;
- holding an annual LT conference with the participation of LT researchers, companies and the public;
- supporting the growth of Icelandic language technology in all possible ways.

Over the past four years, researchers connected to the ICLT, who had been involved in most of the projects funded by the LT Program, have initiated several new projects, which have been partly supported by the Icelandic Research Fund and the Icelandic Technical Development Fund. The most important projects are: IceTagger, a linguistic rule-based tagger (Loftsson, 2007, 2008), IceParser, a shallow parser (Loftsson and Rögnvaldsson, 2007, 2008), Lemmald, a lemmatizer (Ingason et al., 2008) and a context-sensitive spell checker (Ingason et al., 2009). These projects are seen as a contribution to the establishment of a BLARK (Basic Language Resource Kit, cf. Krauwer, 2003) for Icelandic.

The Icelandic LT research group is now in a position to make a research plan for the next few years, building on the resources created and the experience gained in the group's previous work. We know what kinds of resources, tools and methods are most urgently needed, and we believe we know what kind of research needs to be carried out in the near future. We have just received a relatively large Grant of Excellence ("Viable Language Technology beyond English - Icelandic as a test case") from the Icelandic Research Fund to carry out our research plan.

3 Research Plan for Icelandic LT

The existence of LT for any given language could be a deciding factor in whether that language survives the 21st century. The problem is that language resources like treebanks and wordnets are expensive to build and as the corresponding resources for English and other dominant languages become more advanced, the gap between the minority language and the "state of the art" grows. And as English continues to lead the field onwards, even the other dominant languages could struggle to keep up.

Languages other than English face two main problems in LT:

- They have less resources than English to develop LT modules (people and money);
- They may differ from English in important linguistic ways (morphology, syntax, etc.) and therefore the established methods from English LT need adaptation.

Solutions and innovations which address these two problems form the foundation of viable LT for all languages other than English. Although the first problem is a general one, it is particular-

ly acute for languages with small speech communities, such as Icelandic or Faroese, and languages spoken only in countries where economic conditions are unfavourable, such as various African languages. The second problem is moderate or acute depending on the typological distance from English. For instance, English has only sparse morphological inflection and established solutions therefore largely ignore this linguistic property; however, many languages (like Icelandic) have an extremely rich morphology which poses special challenges.

The second problem also relates to how linguistic knowledge is generally harnessed in LT. The rise and success of statistical methods have made the field look like just a branch of applied machine learning in recent years. However, much of the difference between proposed systems lies in the selection of features fed into the machine – but selecting a good feature set is about good linguistics, not good statistics. The tradition in the literature of opposing data-driven statistical methods to hand-crafted linguistic rule methods could therefore be both misleading and harmful (cf. also Trosterud, 2008).

To address the problems for LT viability discussed above, it is essential to develop new methods for constructing LT modules, such as treebanks and semantic databases, in more efficient ways. Our primary objective is to make it realistic to develop three particular types of LT modules with limited resources without sacrificing the quality of the work. The three types of modules are a **database of semantic relations** (Nikulásdóttir and Whelpton, 2009), a **shallow transfer machine translation system**, and a **pilot treebank**. These modules are chosen because they are central to current LT work and prerequisites for further research and development in Icelandic LT. The project will emphasize the following points:

- Developing methodologies for creating resources for new languages more efficiently, with focus on semi-automatic/machine assisted resource generation;
- An inquiry into linguistic issues that are of little relevance for English LT but crucial for many other languages, with a special focus on general methods to deal with morphological richness and morphological ambiguity;
- A case study of Icelandic where we use the tools and methods developed to build a

treebank, a database of semantic relations and a machine translation system;

- Evaluation of the tools and methods developed – focusing on quality of output as well as the output/manpower ratio;
- Writing and publishing guidelines for creating similar LT modules for less-resourced and/or morphologically rich languages;
- Enhancing research training in the field by giving graduate students the opportunity to work on research projects, as it is vital for the future of Icelandic LT to educate and train young researchers in the field.

In short, the project emphasizes the development of viable research methods and practical solutions that will strengthen Icelandic LT and serve as a model for other less-resourced languages.

4 The Prospects of Icelandic LT

The Language Technology Committee estimated that it would cost around one billion Icelandic krónas (then about ten million Euros), to make Icelandic language technology self-sustained (Ólafsson et al., 1999). After that, the free market should be able to take over, since it would have access to public resources that would have been created for money from the Language Technology Program, and that would be made available on an equal basis to everyone who was going to use these resources in their commercial products.

Even though the Language Technology Program was very successful and had a great impact on the development of Icelandic language technology, the fact remains that its total budget over the lifespan of the program (2000-2004) was only 133 million Icelandic krónas – that is, 1/8 of the sum that the committee estimated would be needed. Since then, the LT group has received a number of research grants which amount to approximately 15 million Icelandic krónas. It should therefore come as no surprise that we still have a long way to go.

There are only 300,000 people speaking Icelandic, and that is not enough to sustain costly development of new products. If Icelandic is to survive as a viable national language in the developed world, it must be able to meet IT demands. Consequently, investment in language technology must form an essential part of its language preservation policy. Furthermore, continued public support for Icelandic language tech-

nology will guarantee exploitation of the tools already developed and the knowledge and experience of researchers and companies which has already been accrued. A further way to do this would be to make more use of free/open source licenses, both for software and linguistic resources. It has recently been argued convincingly by several authors (cf., for instance, Forcada, 2006) that it is essential for minor/non-central/less-resourced languages to adopt open source policy with respect to LT resources in order to survive the Information Age.

Unfortunately, many Icelandic resources such as dictionaries and corpora are privately owned, either by commercial companies or individual authors or researchers, and it can be difficult and expensive, or even impossible, to get permission to use them even for research, not to mention for commercial applications. All grants from the Language Technology program were given with the condition that the resources developed would be accessible for anyone wanting to use them in language technology products. However, these resources are not distributed under an open source license and most of them are not free. Even though the license to use them is usually not very expensive, the license fee acts as a barrier for the use of these resources in LT research and development. It would obviously be beneficial for the future of Icelandic LT to implement open source policy, and this has recently been strongly advocated (Trosterud, 2008; Gíslason, 2008).

In our project, we adhere to the recent open source policy of the Icelandic Government. The source code of our research results will be available under different licenses dependant on their intended usage. Most of it will probably be freely available for the development of Open Source software under the GNU General Public License versions 2 and 3 (<http://www.gnu.org/licenses/#GPL>). In accordance with our general policy, the source code of the main programs that we have developed, IceParser, IceTagger, and Lemmald (cf. Section 2) will be made open source in the course of the next few months.

5 Proposals for Nordic Cooperation

Since 2000, Icelandic researchers and policy makers have taken an active part in Nordic cooperation on language technology. This has been of major importance in establishing the field in Iceland. For a small language community and a small research environment like the Icelandic

one, cooperation on LT education, research, use of infrastructures, etc., is vital. The Nordic Language Technology Research Programme 2000-2004 (Holmboe, 2005) was very important in this respect and the continuation of that program or a similar one is absolutely essential.

Some of the smaller language communities in the Nordic/Baltic area still do not have even the most basic LT modules and resources. It is just as expensive to build these modules and resources for the small language communities as for the larger ones, and enough national funding for such development may not be available. For fruitful cooperation involving all the languages in question to be possible, it is necessary to create some minimal common ground, and that means that the smaller language communities need some external support in the beginning. This support can be in the form of direct funding from Nordic funds or programs, but it can also involve exchange of research and knowledge – which then, of course, must be easily accessible.

From 2001-2004, the Nordic Language Technology Research Programme funded language technology Documentation Centers in the five Nordic countries and their cooperation network (NorDokNet; Fersøe, 2005). One of the main goals of the Centers was to collect information on people, projects, products, materials, companies, organizations, etc. having to do with LT in the Nordic countries. Unfortunately, the Centers are no longer funded, and although their web pages still exist, they are not updated as regularly as one would wish and their common website, which has moved to www.cst.dk/nordoknet, is not updated at all.

In 2005, the Nordic Council of Ministers commissioned a ten-year plan in the form of an expert panel report for making the Nordic Countries a leading region in LT (Lindén et al., 2006). One of the main recommendations of the report was the compilation of BLARK reports for the Nordic languages and subsequent funding of LT tools and resources to fill the gaps revealed by the reports. We believe that it would be extremely beneficial to enhanced cooperation to have a common website containing accessible and standardized information on available language resources and tools for the Nordic languages. This could be in the form of a simple table (perhaps on a wiki page for anyone to fill out) with lines for the tools and resources (POS tagger, lemmatizer, monolingual corpus, dictionary, etc.) and columns for the languages. Much of this information can be found on the web pages of the

Nordic Documentation Centers but it does not have a common format, it takes time to collect it, and sometimes it is outdated.

Another aspect of cooperation is education. In 2002, the University of Iceland launched an interdisciplinary Master's program in LT. This is now a joint program between the Department of Icelandic at the University of Iceland and the School of Computer Science at Reykjavik University. The students in the program have had the opportunity to take courses in the Nordic Graduate School of Language Technology (NGSLT). Participation in NGSLT has been absolutely crucial for the Icelandic universities, since they do not have the capacity to give the students high-quality education in LT at home. Unfortunately, the funding period of the school has expired, so this opportunity will not be available after this academic year. It is unclear whether and how we will be able to continue our Master's Program without the availability of the NGSLT courses.

A Nordic Summer School in LT where graduate students and researchers could meet, exchange ideas, attend practical training sessions and pass on technical skills would be very effective in disseminating knowledge and encouraging mutual awareness of ongoing projects, especially if a small number of inspiring international experts were invited to participate in events.

We need to increase and emphasize cooperation in LT teaching and research training – both cooperation between universities and countries, and also cooperation between different fields such as linguistics, computer science, statistics, etc. There have been proposals to start a common Nordic Master's Program but due to lack of funding, it has not been possible to put them into action. It is essential for Nordic LT to find some ways to continue cooperation in this area.

Although both the ICLT and the Linguistic Institute of the University of Iceland are members of CLARIN, Iceland is unfortunately not a member of the CLARIN consortium and thus does not get any funding from the project. Due to lack of domestic resources, Icelandic members have therefore been unable to participate in CLARIN activities. Iceland would obviously have much to gain from the ongoing and planned cooperation within CLARIN, but as things stand, it does not look as if we will be able to take active part in this cooperation in the foreseeable future. It must be a priority task for us to find ways to change this.

6 Conclusion

In this paper, we have demonstrated how joined efforts of the government, research communities, and commercial companies, enhanced by Nordic cooperation, have succeeded in establishing the basis for Icelandic language technology in a relatively short time. We have also outlined the research plan of the Icelandic LT community for the next few years. In addition to its contribution to the building of an Icelandic BLARK, the project aims at developing low-cost methods for building language resources for less-resourced languages. In this respect, we emphasize the importance of open source policy for language resources. Finally, we discuss some ideas for Nordic cooperation on Language Technology, especially as regards compilation and dissemination of information and on LT teaching.

References

- Ari Arnalds. 2004. Language Technology in Iceland. In Henrik Holmboe (Ed.), *Nordisk Sprogteknologi. Årbog 2003*. Museum Tusulanums Forlag, University of Copenhagen, Denmark, pp. 41-43.
- Kristín Bjarnadóttir. 2004. Beygingarlýsing íslensks nútímamáls. [Morphological Description of Modern Icelandic.] In *Samspil tungu og tækni*. Ministry of Education, Science and Culture, Reykjavik, Iceland, pp. 23-25.
- Kristín Bjarnadóttir. 2005. Modern Icelandic Inflections. In Henrik Holmboe (Ed.), *Nordisk Sprogteknologi. Årbog 2005*. Museum Tusulanums Forlag, University of Copenhagen, Denmark, pp. 49-50.
- Hanne Fersøe. 2005. Network of Nordic Language Technology Documentation Centres (NorDokNet). In Henrik Holmboe (Ed.), *Nordisk Sprogteknologi. Årbog 2004*. Museum Tusulanums Forlag, University of Copenhagen, Denmark, pp. 13-16.
- Mikel L. Forcada. 2006. Open Source Machine Translation: an Opportunity for Minor Languages. *LREC-2006: Fifth International Conference on Language Resources and Evaluation. 5th SALTML Workshop on Minority Languages: Strategies for Developing Machine Translation for Minority Languages*, Genoa, Italy, May 23.
- Hjálmar Gíslason. 2008. Gögn og gaman: jarðvegur nýþróunar í tungutækni [The Ground for Innovation in Language Technology]. Paper presented at the workshop *A íslenska sér framtíð innan upplýsingatækninnar?* [Does Icelandic Have a Future within Information Technology?], Reykjavik, Iceland, March 7.

- Sigrún Helgadóttir. 2004. Mörkuð íslensk málheild. [A Tagged Icelandic Corpus]. In *Samspil tungu og tækni*. Ministry of Education, Science and Culture, Reykjavík, Iceland, pp. 67-71.
- Sigrún Helgadóttir. 2005. Testing Data-Driven Learning Algorithms for PoS Tagging of Icelandic. In Henrik Holmboe (Ed.), *Nordisk Sprogteknologi. Árbog 2004*. Museum Tusulanums Forlag, University of Copenhagen, Denmark, pp. 257-265.
- Sigrún Helgadóttir. 2007. Mörkun íslensks texta. [Tagging Icelandic Text.] *Orð og tunga* 9, pp. 75-107.
- Henrik Holmboe. 2005. *Nordisk sprogteknologisk forskningsprogram 2000-2004. Epilog*. NordForsk, Oslo, Norway.
- Anton Karl Ingason, Sigrún Helgadóttir, Hrafn Loftsson, and Eiríkur Rögnvaldsson. 2008. A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI). In Bengt Nordström and Aarne Ranta (Eds.), *Advances in Natural Language Processing*. Lecture Notes in Computer Science, Vol. 5221. Springer, Berlin, Germany, pp. 205-216.
- Anton Karl Ingason, Skúli Bernhard Jóhannsson, Eiríkur Rögnvaldsson, Sigrún Helgadóttir, and Hrafn Loftsson. 2009. Context-Sensitive Spelling Correction and Rich Morphology. *Proceedings of NODALIDA 17*.
- Steven Krauwer. 2003. The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. In *Proceedings of SPECOM 2003*, Moscow, Russia, pp. 8-15.
- Krister Lindén, Kimmo Koskenniemi, and Torbjørn Nordgård (Eds.). 2006. *The Nordic Countries - A Leading Region in Language Technology*. <http://forums.csc.fi/kitwiki/pilot/view/Main/LTExpertPanelReport>.
- Hrafn Loftsson. 2007. Tagging and Parsing Icelandic Text. Doctoral dissertation, Department of Computer Science, University of Sheffield, UK.
- Hrafn Loftsson. 2008. Tagging Icelandic text: A linguistic rule-based approach. *Nordic Journal of Linguistics*, 31(1), 47-72.
- Hrafn Loftsson and Eiríkur Rögnvaldsson. 2007. Ice-Parser: An Incremental Finite-State Parser for Icelandic. In Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek, and Mare Koit (Eds.), *Proceedings of the 16th Nordic Conference of Computational Linguistics*. Tartu, Estonia, pp. 128-135.
- Hrafn Loftsson and Eiríkur Rögnvaldsson. 2008. Linguistic Richness and Technical Aspects of an Incremental Finite-State Parser. *Partial Parsing 2008. Between Chunking and Deep Parsing*. LREC 2008 Workshop, Marrakech, Morocco, pp. 1-6.
- Anna Björk Nikulásdóttir and Matthew Whelpton. 2009. Automatic Extraction of Semantic Relations for Less-Resourced Languages. In *Proceedings of WordNets and other Lexical Semantic Resources – between Lexical Semantics, Lexicography, Terminology and Formal Ontologies. Workshop at NODALIDA 17*.
- Rögnvaldur Ólafsson. 2004. Tungutækniverkefni menntamálaráðuneytisins [The Language Technology Program of the Ministry of Education, Science and Culture]. In *Samspil tungu og tækni*. Ministry of Education, Science and Culture, Reykjavík, Iceland, pp. 7-13.
- Rögnvaldur Ólafsson, Eiríkur Rögnvaldsson, and Þorgeir Sigurðsson. 1999. *Tungutækni. Skýrsla starfshóps*. [Language Technology. Report of a Committee]. Ministry of Education, Science and Culture, Reykjavík, Iceland.
- Eiríkur Rögnvaldsson. 2004. The Icelandic Speech Recognition Project *Hjal*. In Henrik Holmboe (Ed.), *Nordisk Sprogteknologi. Árbog 2003*. Museum Tusulanums Forlag, University of Copenhagen, Denmark, pp. 239-242.
- Eiríkur Rögnvaldsson. 2005. Icelandic Documentation Center for Language Technology. In Henrik Holmboe (Ed.), *Nordisk Sprogteknologi. Árbog 2004*. Museum Tusulanums Forlag, University of Copenhagen, Denmark, pp. 31-33.
- Eiríkur Rögnvaldsson. 2008. Icelandic Language Technology Ten Years Later. *Collaboration: Interoperability between People in the Creation of Language Resources for Less-resourced Languages*. SALT MIL Workshop, LREC 2008, Marrakech, Morocco, pp. 1-5.
- Eiríkur Rögnvaldsson, Björn Kristinsson, and Sæmundur Þorsteinsson. 2006. Nýr íslenskur þulur að koma á markað. [A New Icelandic Text-to-Speech System.] *Morgunblaðið*, January 20th.
- Friðrik Skúlason. 2004. Endurbætt tillögugerðar- og orðskiptiforrit Púka. [Improved Suggestions and Hyphenations in the Púki Spell Checker]. In *Samspil tungu og tækni*. Ministry of Education, Science and Culture, Reykjavík, Iceland, pp. 29-31.
- Trond Trosterud. 2008. Grammar-based Language Technology as an Answer to the Challenges Facing Icelandic and other Circumpolar Languages. Paper presented at the workshop *Á íslenska sér framtíð innan upplýsingatækninnar?* [Does Icelandic Have a Future within Information Technology?], Reykjavík, Iceland, March 7.
- Helga Waage. 2004. Hjal – gerð íslensks stakorðagreinis. [The Making of an Icelandic Isolated Word Recognizer.] In *Samspil tungu og tækni*. Ministry of Education, Science and Culture, Reykjavík, Iceland, pp. 47-53.

CLARIN in Latvia: current situation and future perspectives

Inguna Skadiņa

Institute of Mathematics and Computer Science, University of Latvia
Riga, Latvia

Inguna.Skadina@lumii.lv

Abstract

Although Latvia is a CLARIN member supported only by the government of Latvia, it actively participates in CLARIN project activities. The paper presents current situation in Latvia – existing infrastructure (both LRT and technical), activities taken until now and further work and possible co-operation with NEALT countries.

1 Introduction

Language technologies in Latvia have a rather long history starting at the end of the 50s. While work from the earlier period (the 50s – mid-80s) is fixed now only in research papers, resources and tools developed since the mid-80s are collected carefully and many of them are available on the Web.

Currently the Institute of Mathematics and Computer Science (IMCS) of University of Latvia is the main research institution developing language tools for Latvian, while linguistic resources are created, maintained and preserved in many research organizations in Latvia, including IMCS, National Library of Latvia, universities, many research institutes and some enterprises.

Although the CLARIN initiative has been started only recently, the IMCS has been contributing to CLARIN aims already before by collecting, preserving and making public available linguistic resources, by development the Latvian language tools, by co-operating with other research organizations in resource creation and by being Web publisher and maintainer of resources created in other research institutions.

1.1 IMCS: development of resources and tools

Since 1987 the Artificial Intelligence Laboratory (AILab) at the IMCS of the University of Latvia has been concerned with natural language processing. It is one of the major centres dealing with the collection and exploration of Latvian lexical data in the NLP. Within national (funded by National research programmes “Letonika” and “Informatics”, Latvian Science foundation, structural funds, Latvian State Language Agency, State Culture Capital Foundation) and international projects, different types of data have been collected, analysed and maintained at the Laboratory (Milčonoka et al., 2004; Grūzītis et al., 2004). Many resources are available on the Web (www.ailab.lv) and are used in humanities research since their creation.

Collecting of Latvian resources at the AILab has been initiated at the end of the 80s, beginning of the 90s when fragments of ‘Latvian traditional beliefs’ and some chapters of the first Bible translation carried out in the 17th century by Ernest Glück were keyboarded (Spektors and Baltiņa, 1994). Corpus covering the early written Latvian texts (www.ailab.lv/senie) now contains more than 1 000 000 running words, these are mainly religious texts of the 16th and 17th century (Andronova, 2007).

The collection of modern Latvian texts comprises data from a fiction, a popular science, and some newspapers. At the moment, the number of running words is about 30 millions; about 20 millions words are with HTML mark-up and some 2.5 millions words are with SGML mark-up. A tiny part of data has been morphologically annotated and disambiguated. Recently the balanced corpus of 1 million words (www.korpuss.lv) has been created by support of State Language Agency.

AILab has collected numerous Latvian dictionaries – mainly explanatory dictionaries and dictionaries of terminology. Main resources are: a Term Bank, covering c. a. 115 000 Latvian terms with their translation equivalents into Russian, English, German and Latin (mainly for terms of medicine and biology) and with term definitions where available; Latvian explanatory dictionary; bilingual Latvian-Russian dictionary, electronic version of Mülenbach-Endzelin's 'Let-tisch-deutsches Wörterbuch' (www.ailab.lv/mev), covering c. a. 75 000 headwords and a rich range of examples. The AILab has started an initiative to develop a new electronic dictionary to cover as much Latvian words and their meanings as possible.

Data of a spoken language have been collected and processed at the Laboratory. The speech corpus covers about 20 hours long marked texts.

Apart from lexical data, there are several tools developed for the processing of Latvian: the morphological analyser of Latvian, syntactical analyser, annotation tools etc.

1.2 IMCS: co-operation with other research institutions

National research program *Letonika* aims to facilitate and enhance research activities related to Latvia and the Latvian language, history, culture and other issues. Researchers from sixteen research institutions of Latvia, including the IMCS, participate in this program. Several important resources, such as a database of recently borrowed words, have been created.

Next to national research program *Letonika* many bilateral research projects related to linguistic resources and tools have been realized in co-operation with the University of Latvia and the National Library of Latvia.

Another way of co-operation is storing, maintenance and providing Web access to linguistic resources from the IMCS servers. Currently this type of co-operation has been established with the Latvian Language Institute, the Institute of Literature, Folklore and Arts and the University of Latvia.

2 CLARIN activities in Latvia

In 2006 IMCS and Tilde company were invited to join CLARIN initiative. Since then we actively participate in CLARIN project activities and coordinate CLARIN related activities in Latvia. Latvia is not a CLARIN consortium member yet, however we plan to join CLARIN consor-

tium soon. Until now activities of the IMCS are financed by the Ministry of Education and Science of the Republic of Latvia.

The first year of the CLARIN project was very important for activities in Latvia. This year, two significant activities have been initiated, i. e., the CLARIN Latvia project and the National Corpus of Latvia. Since these initiatives are closely related and the target audience is very similar, we have joined our efforts in dissemination activities and in tasks related to assessing the current state of the art in language resources and tools.

2.1 CLARIN presented to the Latvian State Language Commission

On April 2, 2008 the CLARIN initiative was presented at the workshop organized by the Latvian State Language Commission. It was the first time the CLARIN initiative was presented to the researcher community in Latvia and it received a positive feedback from the participants.

The current situation in language technologies and resources was presented and discussed in a workshop. This workshop gathered ca. 30 participants – representatives of research institutes, universities, publishing houses, libraries and companies dealing with the Latvian language resources.

2.2 CLARIN National Contact Point

One of the first activities in Latvia was establishment of the National Contact Point and development of the CLARIN Latvia Web page (www.clarin.lv). The Web page is used very actively to promote CLARIN related activities in Latvia and Europe. Not only information related to project activities is published, but also different materials which could be useful for users of the infrastructure are published.

Potential contributors and users of CLARIN infrastructure are regularly informed about the CLARIN activities in Latvia by e-mails.

2.3 National seminar

On November 3, a seminar *CLARIN project and the National Corpus* was organized by the IMCS, the Latvian State Language Commission and the National Library of Latvia in order to bring together the potential CLARIN community of Latvia – owners and developers of resources, language technology developers and users of linguistic resources and tools.

The morning session was devoted to the CLARIN project. The CLARIN project coordi-

nator Steven Krauwer presented the mission and role of the CLARIN initiative, emphasizing that all languages (widely and less widely used) are equally important in CLARIN. Participants of the seminar were introduced with CLARIN aims and tasks in Latvia, they were asked to participate actively in the creation of the CLARIN network of expertise in Latvia.

The afternoon session was devoted to the Latvian National Corpus initiative. The current state of Latvian corpus, aims of the National Corpus initiative group, experience of the Czech National Corpus and on-going work on Latvian National Digital Library (being the biggest repository of Latvian culture) were presented in the session.

The meeting was closed by a very interesting discussion on issues related to corpus, copyright issues and access to language tools.

2.4 CLARIN National Advisory Board

The CLARIN National Advisory Board was established during the National seminar with the aim to prioritize goals and tasks of the CLARIN project in Latvia and to facilitate the development of the CLARIN infrastructure.

The CLARIN National Advisory Board includes 17 members from the fields of academia, industry and government. The board members are experts in different CLARIN related issues, such as creation, maintenance and preservation of language resources, development of language technologies, language policy related issues and usage of LRT in social sciences and humanities (SSH). Tasks of the Advisory board include setting priorities and providing recommendations related to goals of CLARIN project in Latvia.

2.5 Workshop on corpus resources

On 4–5 February, a workshop *Corpus of Modern Latvian and its usage* was run at the IMCS. The aim of the workshop was to introduce SSH researchers with possibilities of corpus and corpus exploration tools in their research work. Initially the workshop was planned as a one-day event, but because of great interest, it turned into a two-day session.

The workshop revealed two important issues:

- It is very important to organize practical workshops, where researchers are introduced with possibilities of LRT infrastructure

- There is a big gap between technology and resource providers and users of language resources and tools

This was the first practical workshop; we plan to continue this work by organizing more workshops of this kind.

2.6 Workshop at Rēzekne Higher Education Institution

Rēzekne Higher Education Institution was established only in 1993. Most of teachers are young and open to new technologies and new methods in their research and education. The workshop in Rēzekne was inspired by the corpus workshop in Riga. Participants of the workshop were teachers and students of this institution. Similarly to corpus workshop in Riga this workshop revealed a great necessity for hands-on sessions and discussions on practical issues.

3 CLARIN infrastructure: the state-of-the-art

The IMCS actively participates in the following CLARIN activities: WP2 Technical Infrastructure, WP3 Humanities projects, WP5 LRT Overview, WP8 Construction and Exploitation Agreement. In Latvia work is organized around these activities. Regular internal meetings are held to exchange information between participants from different work packages.

3.1 Overview of LRT in Latvia

The IMCS actively participates in WP5 by collecting and analyzing information about tools and resources developed in Latvia which could serve as a basis for the CLARIN research infrastructure. During the National seminar two questionnaires have been distributed – one concerning resources and tools created in institutions (WP5) and the other about research projects related to the usage of language resources and technologies (WP3).

The results obtained from the questionnaire showed that there are only two institutions, namely, the IMCS and Tilde, who are developers of the Latvian language tools, technologies and applications.

The situation is much better with resources – almost all institutions whose work is related to resources – either they are linguists or computer scientists – have developed or collected some resources. There are many resources in electronic form available, however many of them are available only internally as text files. At the same

time most of the resource owners are interested to share their resources and to include them in the CLARIN infrastructure.

The questionnaire as well as a workshop on corpus usage revealed one problem – even if the resources are publicly available, many potential users don't know about their existence or don't know how to explore or apply them to their own research.

3.2 Technical infrastructure

The IMCS has long term experience in telecommunications and Internet technologies. In 1992 IMCS UL has founded the Academic Network LATNET, in 2007 it was renamed to SigmaNet, the National Research and Education Network (NREN), which provides access to the GEANT2 infrastructure and offers various services.

The main goal of the research is to provide Latvian academic institutions with high quality network services according to the position of the European Union. Research focuses on practical aspects such as design and development of optical networks and deployment of high-performance gigabit network connectivity; data privacy and network security issues; technical and legal aspects of creating and keeping e-documents; legal aspects of networks usage; Grid solutions, methods and software; establishment of Grid infrastructure and the National Grid Centre, etc.

The IMCS has initiated the consolidation of academic Grid resources into the National grid network of Latvia. The IMCS currently has two operational Grid clusters of 12 and 20 CPUs. These clusters are accredited in the EGEE.

On 1 May 2008, the BalticGrid Second Phase (BalticGrid-II) project has started. It is designed to increase the impact, adoption and reach, and to further improve the support of services and users of the recently created e-Infrastructure in the Baltic States.

The IMCS is also participating in the GEANT project and is both a regional NREN (National Research and Education Network) and a CA (Certification Authority) accredited by the EUGridPMA, who coordinates the trust fabric for e-Science grid authentication in Europe.

The institute has experience in execution of large scale NLP tasks in the BalticGRID infrastructure, namely dependency chunking and morphological tagging of the whole Latvian web corpus.

Being member of the CLARIN project is a stimulus to make our language resources and tools widely accessible and compliant with established standards. Our long term intention is to become a CLARIN-conformant national-level service and metadata providing centre. For the preparatory phase, however, we have selected some existing resources and tools that can be rather rapidly integrated in the emerging CLARIN infrastructure.

4 Institutions and co-operation

4.1 The Latvian State Language Commission

The Latvian State Language Commission was established in 2002 by the President of Latvia with the aim to analyze the situation of the state language and to design recommendations for strengthening the status of Latvian as the official language.

The State Language Commission activated necessity to develop language technologies and resources; in order to achieve this, a sub-commission on the Latvian language in New Technologies has been established. The tasks defined by the sub-commission, can be grouped into two major categories. First, tasks related to the creation of a scientific basis for the introduction of the Latvian language use in new technologies. Second, a practical work aimed at the introduction and use of Latvian in new technologies, as well as the use of the new technologies in language development. Tasks set by the sub-commission are included in the State Language Policy Basic Guidelines for 2005–2014 (Vasiļjevs, 2008).

4.2 Latvian National Corpus initiative

For a number of years a development of the Latvian corpus has been among key priorities of the language policy of Latvia. Still practical implementation was hindered by a lack of funding, coordination and a limited awareness in the humanities community. To coordinate the activities of different institutions and to raise general awareness the Latvian National Corpus initiative was finally launched and a working group established in 2008.

The Latvian National Corpus (LNC) initiative has linked efforts of the Latvian State Language Commission, the National Library of Latvia, the biggest resource holders and the universities.

The Latvian National Corpus has been envisioned as a Latvian building block in CLARIN's

common language resource infrastructure. It will be an open on-line resource providing access to federated resources from different research institutions and content providers.

4.3 Co-operation with universities and research institutes

The National seminar and corpus workshop revealed that research community of Latvia has a great interest in implementation CLARIN infrastructure and they are interested and ready to contribute to it.

The following institution expressed their interest in the CLARIN infrastructure: the Academy of Science of Latvia, Daugavpils University, the Institute of Latvian language, the Institute of Literature, Folklore and Arts, the Latvian State Language Agency, the Latvian State Language Commission, Liepāja University, the Ministry of Education and Sciences, the National Library of Latvia, Rēzekne Higher Education Institution, Riga Teacher Training and Educational Management Academy, Tilde, the Translation and Terminology Centre, the University of Latvia, Ventspils University College.

5 Future perspectives

Now, when the potential contributors and users of the CLARIN infrastructure have been introduced to the project, the IMCS continues work on fulfilling the aims of the CLARIN preparation phase.

There are several activities where we see our role in next years of the preparation phase. First, we will continue to contribute to the EU CLARIN project and will work to prepare Latvia for the construction phase of the project.

Second, we will continue activities related to knowledge transfer to SSH research community. Already after corpus seminar several institutions showed interest to contribute their resources. We will continue to provide technical support to them.

Third, we plan to implement the CLARIN infrastructure prototype at least for the IMCS, thus becoming a real CLARIN centre.

Fourth, we will actively co-operate with other institutions in Latvia to create nationally important resources, such as the National Corpus.

Until now co-operation with other CLARIN countries was based on informal discussions of the CLARIN implementation scenarios in other countries, however we are open for closer co-

operation in future, especially with Baltic region and NEALT countries.

Acknowledgements

The CLARIN project activities are supported by the Ministry of Education and Science of the Republic of Latvia within project 'Participation of Latvia in the first period of preparation phase of the CLARIN project' (September, 2008 – April, 2009). Author would like to thank my colleagues Everita Andronova, Normunds Grūzītis, Ināra Opmane and Andrejs Spektors from IMCS and Andrejs Vasiļjevs from Tilde who contributed to this paper.

References

- Andronova, Everita. 2007. The Corpus of Early Written Latvian: current state and future tasks. *Proceedings of Corpus Linguistics 2007*, Birmingham, UK. Electronic publication: (http://ucrel.lancs.ac.uk/publications/CL2007/paper/245_Paper.pdf).
- Grūzītis, Normunds., Ilze Auziņa, Sanita Bērziņa-Reinsone, Kristīne Levāne-Petrova, Everita Milčonoka, Gunta Nešpore, Andrejs Spektors. 2004. Demonstration of resources and applications at the Artificial Intelligence Laboratory, IMCS, UL. *Proceedings of the first Baltic conference 'Human Language Technologies – the Baltic Perspective'*, Riga, pp. 38–42.
- Milčonoka, Everita, Normunds Grūzītis, Andrejs Spektors. 2004. Natural language processing at the Institute of mathematics and computer science: 10 years later. *Proceedings of the first Baltic conference "Human Language Technologies - the Baltic Perspective"*, Riga, pp. 6–11.
- Spektors, Andrejs and Maija Baltiņa. 1994. Latviešu valodas vēsturisko tekstu datu bāzes izveide. *Valoda un tehnika Eiropā 2000*, Riga, p. 30.
- Vasiļjevs Andrejs. 2008. The influence of new technologies upon the Latvian language. *Break-out of Latvian*. Zinātne, pp. 345–355.

The Possible NEALT Role in the Consolidation of the Nordic and Baltic Language Resources

Pavel Skrelin

Saint-Petersburg

State University

Saint-Petersburg, Russia

skrelin@phonetics.pu.ru

Vera Evdokimova

Saint-Petersburg

State University

Saint-Petersburg, Russia

postmaster@phonetics.pu.ru

Karina Evgrafova

Saint-Petersburg

State University

Saint-Petersburg, Russia

evgrafova@phonetics.pu.ru

Abstract

This paper discusses the issues of possible cooperation among different countries within the NEALT Geographic Region for constructing an infrastructure of common language resources in connection to the European CLARIN (Common Language Resources and Technology Infrastructure) initiative. The information about the national projects in language technology area in North-West Russia (Saint-Petersburg) is presented. It is suggested to discuss the possibility of sharing speech and language resources and special tools for different national speech and text corpora elaborated in the NEALT-countries.

1 Introduction

The cooperation between the Nordic countries within NEALT seems to be fruitful. It can provide different opportunities of sharing approaches to the scientific problems in language technology studies. The cooperation in teaching helps to exchange knowledge in different fields of language science. The opportunity for language technology (LT) Master students and PhD students to travel and study around the Nordic and Baltic region is to make them better specialists who can deal with the language problems and create common tools for various languages. Therefore, it appears to be useful to share LT resources and standards for speech and text corpora descriptions. The compatible formats and tools for working with language databases can be a very good result of NEALT-cooperation.

2 National work in LT

2.1 Previous work in LT studies in NW Russia

First, we would like to dwell on the situation with LT studies and LT science at the Department of Phonetics at Saint-Petersburg State University in Russia [1].

The research in the language and speech technology emerged in the 1950ies in Russia and was established in the early 1960ies at various academic institutions. In the 60-ties the All-Russia workshop in "the Automatic recognition of sound patterns" was started and it existed until 1990. Thus the research teams from academic institutions and universities had a good opportunity to discuss problems in the LT and speech technology (ST) domain every two years in addition to different international and national conferences.

In 1996 a project funded by the national foundation "Integratia" was started. It aimed at bringing together and integrating efforts of leading research teams in Saint-Petersburg which were involved in research into the models of speech-to-speech translation (English-Russian and Russian-English). The project was performed by the Department of Phonetics SPbSU (speech synthesis), the Laboratory of Engineering Linguistics of the Russian State Pedagogical University (machine translation) and the Laboratory of Speech of the St.-Petersburg Institute of Informatics of the Russian Academy of Sciences (speech recognition and understanding). In the framework of this project students from SPbSU and RSPU could freely take courses from the other participating university and from the Institute of Informatics.

The Phonetic Fund of the Russian language is the project which started more than 20 years ago. The Phonetic Fund is conceived and developed as a collection of three related components:

- 1) acoustic material,
- 2) software tools for its processing and analysis,
- 3) the results of this analysis.

The contents of the Fund are a collection of all forms and significant units of the Russian language taking into account all its variants and dialects.

The acoustic databases are designed for the storage of the phonetically representative sound material. A part of the sound material from the Phonetic Fund of the Russian Language is presented in the format, of the phonetically representative text, composed of 200 most frequently occurring Russian syllables in all possible rhythmic positions. The Russian phonetically representative text has been recorded from four Russian speakers (2 male and 2 female speakers) representing Moscow and St. Petersburg pronunciation standards, and also from several foreign speakers (Bulgarian, Finnish, American English, Korean, etc.) that demonstrates phonetic interference.

Sound archives (acoustic databases produced from old sound recordings collections of the Institute of the Russian Literature):

Zhirmunsky's collection" of old recordings of the folklore of so-called "Russian Germans" – Germans who lived in the Volga region since XVI century. The recordings were made in the 20-s and 30-s in Russia.

Another archive is presented by "Tales of the Russian North" and "Poetic Folklore of the Russian North (lamentations)". In the dialects of these outlying regions (Pechora, Arkhangel'sk, etc.) one can find the traces of very ancient states of the Russian language.

2.2 Collaboration in LT in NW Russia

Speech technology as the major subject is only taught at the Department of Phonetics at Saint-Petersburg State University.

These are some of the areas of research and expertise of Department of Phonetics and the Laboratory of Experimental Phonetics:

- automatic text processing: parsing, automatic phonemic and phonetic transcriptions, intonation transcription;
- computer-assisted speech signal analysis and modification;

- speech signal segmentation (including automatic segmentation) into sounds, intonation units, phrases;
- automatic pitch tracking;
- acoustic databases, speech corpora, speech synthesis, speech recognition, computer-assisted language learning programs.

SIGRU

The ISCA Special Interest Group on Russian Speech Analysis (SIGRU) has the overall aim of promoting research and development in the scientific, technical, professional and didactic fields of speech and language technology for Russian speech analysis, particularly formal methods of analysis [2]. The group covers the staff of the Department of Phonetics, researchers from The Laboratory for Experimental Phonetics and researchers specializing in the Russian language from different parts of Russia and other countries.

SIGRU pursues the following purposes.

1. Promoting and organizing conferences, schools and workshops;
2. announcing publications (papers, theses and dissertations) on topics related to Russian speech analysis and/or by authors that are members of this SIG;
3. promoting industry - university collaboration;
4. promoting interdisciplinary scientific communication of researchers dealing with speech analysis;
5. promoting scientific and technical exchange of information;
6. providing a channel of communication between Russian speech researchers and those active in speech and language technology in general.

The Department of Phonetics collaborates with a number of organizations working in LT and ST fields in NW Russia. Such fields of LT as machine translation and text processing are investigated in the Laboratory of Engineering Linguistics of the Russian State Pedagogical University in Saint-Petersburg [3]. The Laboratory of Speech of the St.Petersburg Institute of Informatics of the Russian Academy of Sciences [4] specialises in automatic speech recognition, automatic speech understanding.

Several companies working in the field of LT and ST collaborate with the Department of

Phonetics (f.i. The Speech Technology Centre, Auditech Ltd).

2.3 Russian Spontaneous Speech Corpus

In 2001-2007, the Laboratory of Experimental Phonetics developed a Russian spontaneous speech corpus comprising recorded speech by 10 speakers labeled with boundaries of segmental and prosodic units. This work was conducted as part of the different projects supported by INTAS, RFBR, Ministry of Science and Education and was supported by The President's grant for the leading scientific schools ("The Characteristics of Segmental and Prosodic Units in Different Types of Speech: Standard and Current Trends") [5].

2.4 Speech Database for Russian TTS synthesis

A large speech database has been recorded and is being annotated for unit selection synthesis system.

Each database contains about 10 hours of speech for 8 speakers. Two hours are segmented at different levels manually; the rest of the segmentation is performed automatically in the force alignment mode of a Russian speech recognition system developed at Speech Technology Center. The database contains reading different texts read by the speakers. Some of texts are aimed at obtaining intonation-rich and expressive speech.

3 Cooperation in NEALT-countries

The text and speech corpora and archives mentioned above may be of great interest for comparative studies in the field of LT for the Nordic Languages. However, there does not seem to be an agreement among the linguists about the common standards for speech corpora annotation. Therefore, it could be very interesting to discuss the issues of sharing the speech resources, special tools for conversion and standards within the Nordic countries. It may be done by the joint effort of researchers from the Nordic countries.

3.1 Workshops

Workshops to discuss the possibility of discussing standards for text and speech corpora annotation may be held regularly in the Nordic countries. The goal of these meetings is to take a closer look at the existing speech corpora, to discuss the possibility of sharing tools, methods and approaches for working with them.

These workshops can be also aimed at providing a forum for researchers to present their work in this field and to discuss future developments such as building shared resources and can be held with the conferences or seminars performed by NEALT.

Candidate topics of interest include:

- the structure of different corpora types;
- proposals for annotating corpora;
- tools for annotation conversion for different types of corpora.

3.2 Teaching opportunities

On the whole, language technology teaching in the Nordic and Baltic countries and in NW Russia is on its way to the common Bologna style university degrees of roughly equal measures and equivalent contents. Though the goal is common, the pace in moving to the common system varies as well as the present stage where countries and individual institutions presently are.

In the CLARIN project language technology includes both speech technology and text-based natural language processing and also many of the application areas of the core language technology methods and theories. Language technology is studied and taught in a variety of contexts including linguistics, computer science, information sciences, electrical engineering and other more established disciplines along where the subject is explicitly called language technology or computational linguistics. Being a multidisciplinary subject, language technology may even benefit from this diversity by being able to offer more variety and contacts to related theories and methods.

PhD students from different countries can travel and study in different NEALT-countries. This can help sharing the annotation approaches and resources. The opportunity for LT Master students and PhD students to travel and study around the Nordic and Baltic region is to improve their professional skills.

It can be effective to provide short-term courses and seminars for LT students from different Nordic and Baltic countries. Therefore, it may be useful to include the information about different speech and text resources and special tools for their processing. Thus, the unified approaches to annotation, formats and standards can be devel-

oped by the NEALT-countries. The available databases can be used during the lectures and seminars as an example material.

The NGSLT School has got 5-year successful experience of such studies within the Nordic and Baltic countries (www.ngslt.org).

There are other educational programmes within Europe, such as Erasmus Mundus 2009-2013. It is a cooperation and mobility programme in the field of higher education that aims to enhance the quality of European higher education and to promote dialogue and understanding between people and cultures through cooperation with third countries [6].

4 Conclusions

Thus there seem to be good prospects for possible cooperation among the Nordic countries.

References

- [1] <http://www.phonetics.pu.ru>
- [2] <http://forma.pu.ru/en/index.html>
- [3] <http://www.prikladnaja.narod.ru/>
- [4] <http://www.spiiras.nw.ru/modules.php?name=Content&pa=showpage&pid=85>
- [5] <http://www.speech.pu.ru>
- [6] http://eacea.ec.europa.eu/static/en/mundus/erasmus_mundus_2009_2013_en.htm

CLARIN: Norwegian and Nordic perspectives

Koenraad De Smedt

University of Bergen and Unifob AKSIS

Abstract

This position paper addresses the question whether there is a need for Nordic cooperation on building a language infrastructure for the Humanities, given the existing European cooperation in the CLARIN project and a number of national initiatives. It will be argued that the Nordic level is not superfluous, but in fact it seems the most efficient and appropriate level for cooperation, based on size, common culture, cooperation record and existing frameworks.

1 Status of coordination of language resources in Norway

Not since the Norwegian Computing Centre for the Humanities, established in Bergen in 1972, was discontinued as a centre with national responsibilities in the 1990s, has Norway had a centralized coordination of digital resources for the Humanities. A plethora of activities has taken place in the past four decades, resulting in a wealth of digital resources and technologies, many of very high quality, but as a whole rather disparate and not easy to access for exploitation.

The KUNSTI research program (2001–2007) provided an important stimulus to language technology research in Norway, but this program was not targeted at unifying existing language resources into a usable whole. Since 1999, there have been numerous surveys, studies, reports and plans aimed at building a comprehensive Norwegian language bank, but this goal as such has not received adequate funding so far. A Norwegian government proposition in 2008 stated that the Norwegian HLT Resource Collection shall be established on Jan. 1, 2009, but although this activity has received considerable moral support from the government, the Language Council and the

universities, it has so far not received substantial funding. Besides, the latter initiative is targeted at language technology development and not at a comprehensive language infrastructure for the Humanities.

2 The *Infrastruktur* program

Recently, Norway's strategic investment in research infrastructures was accelerated by initiatives abroad, especially by Europe's engagement in scientific infrastructures since the 2000 Strasbourg Conference on Research Infrastructures, leading to the first ESFRI roadmap in 2006 and the projects in the Capacities program since 2007.

In early 2008, a strategy document was published, *Verktøy for forskning: Nasjonal strategi for forskningsinfrastruktur (2008–2017)*, that envisaged the establishment of a special research fund of NOK 20 billion with a yearly yield of NOK 800 million,¹ 75% of which would be channeled through the Research Council of Norway (RCN) and 25% of through R&D institutions.

Further recommendations, including continued participation in NDGF (Nordic Data Grid Facility), have come through the strategy document *Nasjonal strategi for eInfrastruktur*, which outlined in particular the electronic platforms necessary for digital infrastructures.

The first call for proposals in research infrastructures under the program *Infrastruktur* was published in early 2009 by the Research Council of Norway (RCN), with an initial overall budget framework of approximately NOK 400 million for the initial announcement. This call, with a deadline of April 22, is open to all scientific disciplines and encompasses several categories of research infrastructure described in specific calls, among

¹Actual allotted amounts are dependent on the national budget. The current national budget partly complies this proposition, and a government fund with a start capital of NOK 4 billion will be established.

which the category most relevant to CLARIN is *Scientific databases and collections*. This program currently seems the best option for building up some of the language resources and technologies that will make up Norway's contribution to CLARIN, although heavy competition can be expected from all scientific disciplines, witnessed by the fact that in the pre-proposal round, the call was oversubscribed by a factor of 25.

On the one hand, a number of proposals for rather specific large-scale language resources are being prepared for this call, such as a database for speech and dialect data, one for syntactically and semantically annotated corpora, etc. It is expected that these infrastructure projects, if funded, will strive to be compatible with CLARIN, but it does not currently seem guaranteed that the results will in fact be incorporated in CLARIN. There is not even a plan for joining these resources in a single national infrastructure for language resources.

On the other hand, a coordination project entitled NO-CLARIN is submitted that ensures national networking and liaison of national activities to the CLARIN effort. NO-CLARIN will promote networking between actors and stakeholders in Norway through events and other communication. It will also run case studies and pilots to investigate the possible establishment of a Norwegian CLARIN center, while Nordic cooperation in this area will also remain a possibility. NO-CLARIN builds on previous coordination activities in late 2008, in particular a national seminar with 36 participants which also included several representatives from other Nordic countries. Current support schemes at RCN only cover networking and preparatory activities under the preparatory phase of ESFRI projects, while schemes for national support under the next phase of ESFRI projects are not yet available.

3 Nordic cooperation on language technology

The Nordic countries have a good record of cooperation and mutual understanding, partly thanks to regular cooperations in higher education, researcher training and research projects, partly stimulated by specific programs, in particular the recent Nordic language technology program (2000-2004, extended to 2005). The networking activities stimulated by this program did not only focus on specific research fields, but also in-

cluded a coordinated documentation activity (NorDokNet) and an outreach to the Baltic countries in 2005. An extension and consolidation of these cooperation and networking efforts was attempted through bids for a Nordic Center of Excellence (2005), a Nordic documentation effort with industry through Nordisk InnovationsCenter (2005), and a Joint Nordic Use of Infrastructures (2007), but all three bids were unsuccessful.

However, in 2006 the Northern European Association for Language Technology (NEALT) was founded and established good publication channels. Furthermore, the Nordic language councils have a good tradition of cooperation that also encompasses the stimulation of language technology applications for the Nordic languages. As part of this cooperation, a working group on *Språkvård och språkteknologi i Norden* was established and a report *SpråkVis — Språkteknologisk vismansrapport* was ordered. It is in this spirit of Nordic cooperation and language appreciation that further joint work on language resources and technologies seems feasible.

4 Nordic initiatives in e-infrastructures

The Nordic countries have a number of instruments promoting research cooperation. In particular, The Nordic Council of Ministers provide funding of common actions in education and research through programmes and actions administered by NordForsk. One recent NordForsk initiative is The Nordic eScience Initiative, which may bear relevance to CLARIN. Its task is to "... describe what Nordic level functions and services would be beneficial for coupling digital resources using Grid technology, including computational resources, data repositories and key research instruments. The proposed functions and services should, by federating resources and competences, add value to Nordic research communities and to the NGIs. Furthermore, the proposal may propose a joint Nordic framework for resource provisioning and sharing/aggregating national resources. The Nordic centers/metacenters have already made significant progress in this direction." From this description, it appears that this task could be a good match for reaching the goals of CLARIN at a Nordic level.

5 Perspectives for Nordic cooperation on CLARIN

As mentioned above, there has been an unsuccessful attempt to obtain funding for a Joint Nordic Use of Infrastructures, but with careful planning, joint Nordic activities may still be realized. I believe that Nordic cooperation is beneficial because Nordic projects in this area will have the most efficient dimension. Nordic countries have good expertise, but since research groups are small, it is only through pooling that a critical mass will be reached. On the one hand, even at a national level, the research capacity in the area of language resources and technologies of a country like Norway is quite limited. On the other hand, full interaction between 23 countries at a European level is quite complex and requires enormous management resources. In contrast, Nordic cooperative projects would be of a manageable size, but at the same time they embody a sufficient economy of scale.

Research infrastructures are expensive to establish and run. CLARIN is currently estimated to cost EUR 23.2 million in the construction phase. While the data throughput on the CLARIN grid is expected to be smaller than typical amounts, for instance, in particle physics or climate research, language data is more heterogeneous and structured, such that curation of language data, as well as search in annotated data, is more complicated and expensive than for the huge amounts of data that is produced by the Large Hadron Collider experiments. CLARIN will therefore be a distributed facility relying on networked centers with special expertise at specific centers.

There will be also a need for physical platforms with large media for datastorage and supercomputers that perform searches in databases with good response time.² It is inevitable that demands will be placed on cost-efficiency; such demands are already being made in the Norwegian *Infrastruktur* program. In this context, the benefits of cooperation and the necessity to operate swiftly and efficiently make it natural to consider extending national networking efforts once again to a Nordic level, perhaps in the following ways:

1. Communication forums and meetings ought to be established to exchange and discuss

²Trebank searches, for instance, may involve arbitrarily complex graph traversals that place heavy demands on CPU power and memory.

common experiences, proposals and solutions, for instance through Nordic workshops on language infrastructure research and through invitations of other Nordic partners to national seminars.

2. A liaison ought to be established between the Nordic partners in CLARIN and relevant Nordic actors in e-Infrastructure, including the eNoria Task Force on Sustainable Nordic Grid Collaboration, and NDGF, with the view of exchanging information between linguistic and technical communities.
3. The linking of national language infrastructure centers in a Nordic grid solution ought to be investigated and tried out in case studies and experiments. Such a grid might in the first instance be easier to achieve on a Nordic scale than on a full European scale.
4. Financing possibilities in order to support some of the above actions ought to be looked at on a Nordic level, perhaps also on national and European levels.

6 Conclusion

The main reasons for working on a Nordic level are the following. First, relevant actors at the Nordic level know each other, have a record of cooperation, and share a common culture (including a research culture). Second, there is an important ‘Goldilocks’ argument of finding the right size: whereas research communities in most of the Nordic countries are too small, and the European community may be a bit too big, the Nordic community seems just the right size. Third, there are existing Nordic cooperative initiatives in eScience that may serve as a frame, platform or jumping board, whichever metaphor one prefers. The best thing to hope for is that research and infrastructure activities on the various levels (local, national, regional and European) will not be in the way for each other, but will complement each other in the spirit of *subsidiarity*, in the sense that activities should be managed on the level where it is most efficient to do so.

7 Links

1. <http://www.clarin.eu>
2. <http://www.spraakbanken.uib.no/utredninger.page>

3. <http://www.regjeringen.no/nb/dep/kkd/dok/regpubl/stmeld/2007-2008/stmeld-nr-35-2007-2008-.html?id=519923>
4. <http://cordis.europa.eu/esfri/>
5. <http://link.uib.no/?vhuj>
6. <http://www.rcn.no>
7. <http://www.ndgf.org/>
8. <http://link.uib.no/?JAD3>
9. <http://www.cst.dk/nordoknet/>
10. <http://omilia.uio.no/nealt/>
11. <https://kitwiki.csc.fi/twiki/bin/view/Main/LTExpertPanelReport>
12. <http://www.nordforsk.org>
13. <http://www.nordforsk.org/text.cfm?id=499>