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 recourses 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 untill 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 envolved 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 Pedagogigal University (mashine 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 if 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 fileds of LT as machine translation and text processing are investigated in the Laboratory of Ingeneering Linguistics of the Russian State Pedagogigal University in Sain-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 end 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