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 co-operation. 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 internetional 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

Promt (St. Petersburg), Fotonija (Vilnius), Alna 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 biuras* 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 programs. 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 (Čulo 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 Čulo, Gintarė Grigonytė, Merylyne Hermandez, Algirdas Avižienis, Johann Haller, Rūta Marcinkevičienė. 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 Humantities in Europe*. The METRIS Report.

Rūta Marcinkevičienė. 2006. Baltų kalbų išlikimo problema informacinėje visuomenėje (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.