INFORMATICA, 1998, Vol. 9, No. 4, 449–456

Language Egineering in Lithuania

Joana LIPEIKIENĖ, Antanas LIPEIKA
Institute of Mathematics and Informatics
Akademijos 4, 2600 Vilnius, Lithuania
e-mail: lipeika@ktl.mii.lt

Abstract. Language engineering encompassing natural language processing and speech processing became very important for a development of every nation in multilingual Europe. After the Council of European Union approved conclucions on linguistic and cultural diversity, tools and systems created for every European language are necessary to overcome language barriers and to use all languages in various spheres of human cooperation. The paper gives an overview and a consideration of language engineering in Lithuania.

Key words: language engineering, natural language processing, speech processing.

1. Introduction

The concept Multilingual Information Society became very popular over the last few years. It emerged in 1994 as the title of a draft Communication and Community Programme of the European Comission. The Council of the European Union approved a number of conclucions on linguistic diversity and multilingualism. In particular, the Council affirmed the importance of linguistic diversity in the European Union, emphasizing that it has important implications for democracy, culture, social and economic development [1]. “In the future any tool, service or system will need to accomodate or be adaptable to any language” [2].

The complicated problems of multilingualism are being solved by joined efforts of researchers working on Natural language and Speech processing in all countries. Integration and cooperation of Language and Speech communities in Europe were especially characteristic to the last decades:

EACL (http://issco-www.unige.ch/eacl/eacl.html) – the European Chapter of Association of Computational Linguistics is the scientific and professional society for people working on natural language processing. It was established in 1982.

ESCA (http://ophale.icp.grenet.fr/esca/esca.html) – The European Speech Communication Association, established in 1988, promotes speech communication science and technology in a European context, both in the industrial and academic areas, covering all aspects of speech communication (acoustics, phonetics, natural language processing, artificial intelligence, cognitive science, signal processing, etc.).

ELSNET (http://www.elsnet.org) – the European Network in Language and Speech, established in 1991, connects both natural language and speech organizations in Europe.

ELRA (http://www.icp.grenet.fr/ELRA/home.html) – the European Language Resourses Association, established in 1995, pursue the object to provide a centralized organization for the validation, management and distribution of speech, text and terminology resourses and tools.

According to the common terminological conventions, language engineering encompases natural language and speech processing. The main parts of language engineering can be briefly represented (see Fig. 1).

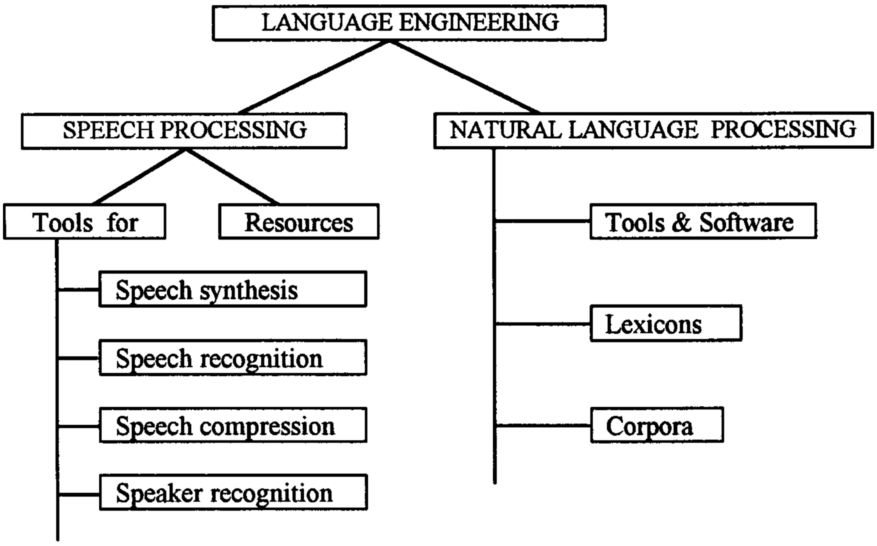

Fig. 1. Main parts of language engineering.

The items of this diagram are the main parts of language engineering that are more or less being developed for every language. The main European languages (English, German, French, Spanish and so on) have tools for speech recognition and synthesis, automatic translation systems, dialogue systems for services, huge language and speech resourses: corpora, multilingual lexicons and terminological data banks etc.

Lithuanian language needs of creating all these tools in order to integrate Lithuanian language among other European languages.

2. Natural Language Processing in Lithuania

What is the situation of language engineering in Lithuania? In the Elsnet Survey of 1994 [3] one can find words "language engineering in Lithuania consists of speech processing only". In 1994 it was true: four main organizations at that time carried out some investigations on speech processing, while language investigations were carried out mostly without use of computers. But the situation has changed during the last few years. Table 1 consists of organizations that are in one or other way connected with language processing.

So there is a number of organizations connected with natural language processing. What are the achievements?

Table 1

The list of language organizations in Lithuania

Name of organization	Main activities developed
Department of General Linguistics, Vilnius University [4]	Investigations of the prosody and spectral features of Lithuanian language
Lithuanian Linguistics Department, Vilnius Pedagogical University [5]	Investigations of Lithuanian prosody by experimental methods
Center of Computational Linguistics, Vytautas Magnus University [6]	Compilation of a big monitor corpus of the written contemporary Lithuanian language
Department of Language history and dialectology, Institute of Lithuanian Language [7]	Research on Lithuanian language history: collection and investigation of dialects, investigation of old written language; compilation of well balanced corpus of Lithuanian language (about 1 mln. words)
Department of Dictionaries, Institute of Lithuanian Language [8]	Compilation of thesaurus type of great "Dictionary of Lithuanian Language"
Institute of Lithuanian Language and Folklore [9]	Research on the development of the Lithuanian artistic language
Laboratory for Computer Applications in Research, Institute of Mathematics and Informatics [10]	Creation and compilation of electronic dictionary of the modern Lithuanian language, frequency dictionary of modern written Lithuanian, terminological data bank including various domains
Philological Department, Šiauliai Pedagogical University [11]	Compilation of alphabetic, frequency and inversive dictionaries of Lithuanian, frequency dictionaries of Lithuanian children literary texts
Recognition Processes Department, Institute of Mathematics and Informatics [12]	Computational morphological analysis and synthesis, multimedia project "Lithuanian dialects" [14]
Computer Software Department, Kaunas University of Technology [13]	Building up and investigation computer based learning systems and courses as well as flexible and distance learning technology
Publishing House TEV [15]	Compilatation of electronic dictionaries for MS DOS, Windows 3.XX, Windows 95, OS/2
Fotonija Ltd. [17]	Production of various Lithuanian language support packages for Windows 3.1X, Windows NT and Windows 95
Sekasoft Ltd. [18]	Development and distribution of computer fonts and tools for Lithuanian language

Fig. 2 consists of items that can be included to the common scheme of natural language processing. Not all of them are completed but all these things that are done or are under construction should form the basis for further investigations of Lithuanian language by the aid of computers.

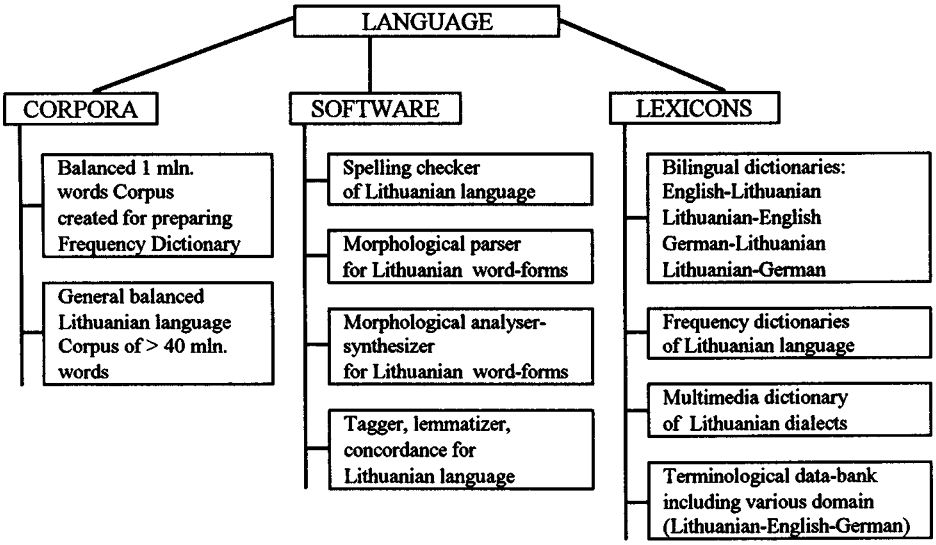
LANGUAGE
CORPORA
SOFTWARE
LEXICONS
Balanced 1 mln. words Corpus created for preparing Frequency Dictionary
General balanced Lithuanian language Corpus of > 40 mln. words
Spelling checker of Lithuanian language
Morphological parser for Lithuanian word-forms
Morphological analyser-synthesizer for Lithuanian word-forms
Tagger, lemmatizer, concordance for Lithuanian language
Bilingual dictionaries: English-Lithuanian Lithuanian-English German-Lithuanian Lithuanian-German
Frequency dictionaries of Lithuanian language
Multimedia dictionary of Lithuanian dialects
Terminological data-bank including various domain (Lithuanian-English-German)

Fig. 2. Presently available Lithuanian language processing tools (some of them are under construction).

One important problem is lithuanization of Internet pages. For presenting and reading of Lithuanian texts on Internet there exists the problem of Lithuanian fonts. Ltd. FOTONIJA and SEKASOFT develope and distribute various Lithuanian language support packages for Window 3.1X, Window NT and Window 95. One can find material on lithuanization of WWW pages and instructions “If your Netscape navigator do not show Lithuanian fonts” in [19], [20].

Our investigations of the situation show very little attention is paid to natural language processing, some of language organizations do not have access to Internet and e-mail or work without computers at all. In Lithuania there is not computational linguistics or related speciality at any university, so one can say people working in this sphere are self-educated. In addition, very small number of people work on natural language processing, in some organizations only one or two. So all three strands of natural language processing are only at the beginning of the work that have to be done.

The general corpus of Lithuanian language is under construction but still it is not of desirable quality and is not available for practical use. As regards the computer based morphological and syntactic analysis of Lithuanian language there are only some attempts to do this. Much more people should do this work in order to automatize analysis of complicated Lithuanian language. For suitable language processing here is a need of much more multilingual lexicons, including various domain multilingual dictionaries.

3. Speech Processing in Lithuania

There are four main organizations working on speech processing. They are enumerated in Table 2.

Table 2
The list of speech organizations in Lithuania

Name of organization	Main activities developed
Speech Research laboratory, Kaunas University of Technology [21]	Phoneme based speech recognition, speech processing, compresion, segmentation
Recognition Processes Department, Institute of Mathematics and Informatics	Speaker recognition
Department of Phonoscopic Examination, Institute of Forensic Examination [22]	Speech analysis, pitch determination for speaker identification
Department of Informatics, Vilnius University	Speech synthesis from Lithuanian texts

Speech processing in Lithuania have been investigated much earlier than natural language processing.

Speech Research Laboratory at Kaunas University has some outstanding achievements in speech recognition, researchers of the Laboratory took part in the development of many international projects and are working on various aspects of speech processing at present. They use modern models of speech (Hidden Markov models, Neural networks, etc.).

Speaker recognition investigations in the Department of Recognition Processes together with the Department of Phonoscopic examination (Institute of Forensic Examination) were successful and are used in forensic examination in practice. But the lack of good resourses – large speech corpora – for testing of investigation is one of the main disadvantages in this work.

Some work on the way to speech synthesis from Lithuanian texts was done at Vilnius University. The small group of people worked on this problem but the work stopped when the people left the University.

So one can summarise a research work on speech processing in Fig. 3.

The research on speech is not coordinated well. Some projects promoted investigations but after some time the work stopped again. The lack of speech resourses and good equipment are the main disadvantages.

It is necessary to consider speech processing and natural language processing together because for successful language engineering the cooperation of speech and language organizations is necessary. For example, it is impossible to create a speech recognition system of good quality without use of specific features of the language. Lithuanian language morphology is characterized by archaic and complicated word – inflecting and word – building system. And it has many other pecularities that make its way to automatic processing complicated. For example, there are many words in Lithuanian that are

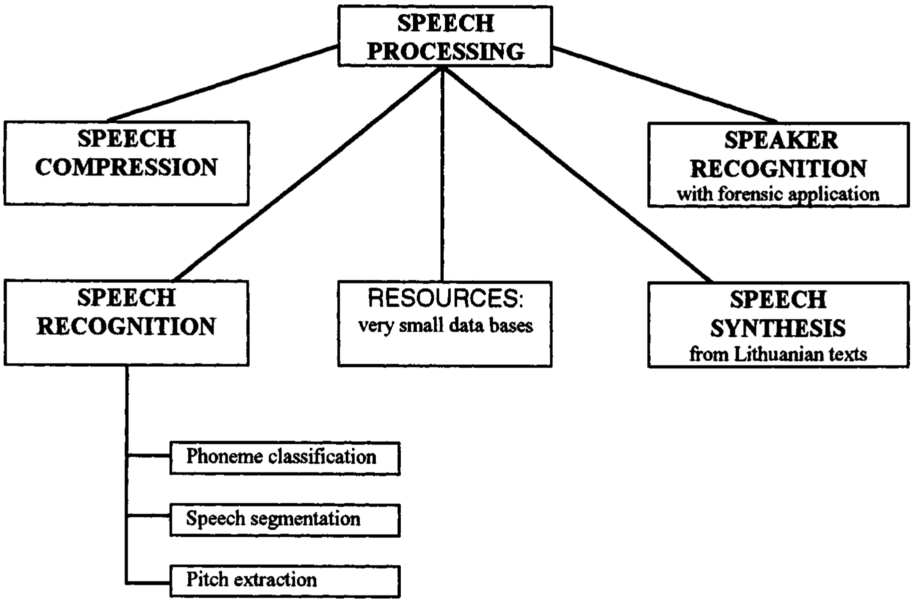
SPEECH PROCESSING
SPEECH COMPRESSION
SPEAKER RECOGNITION with forensic application
SPEECH RECOGNITION
RESOURCES: very small data bases
SPEECH SYNTHESIS from Lithuanian texts
Phoneme classification
Speech segmentation
Pitch extraction

Fig. 3. The main topics of speech processing in Lithuania.

written identically and have the same stressed syllable but because of different accent these words have quite different meanings. On the other hand, people working on natural language processing need knowledge on speech processing for creation of a suitable software for natural language processing. One example of speech and language people cooperation is the project "Lithuanian dialects" [14]. Researchers at the Institute of Lithuanian Language developed this project together with UNESCO Chair in Informatics for the Humanities[16]. It is presented on Internet in multimedia form and one can look at the map of Lithuanian dialects and listen the main Lithuanian dialects.

4. Conclusions

The overview of speech and language research shows that it is necessary to develop all strands of Language engineering in Lithuania.

For natural language processing it is especially important at least

- to make the general Lithuanian language corpus of desired quality and available for practical use in research;
- to improve available and to create other desirable software for computer based morphological and syntactic analysis of Lithuanian language;
- to create more multilingual lexicons, including teminological various domain multilingual dictionaries.

For speech processing one of the most important tasks is to create spoken Lithuanian database in agreements with standards and recomendations applied for the most spoken

European languages corpora. It is necessary for appropriate phonetic research, for training and evaluation of speech, speaker recognition systems and for development of corpus – based speech synthesis systems.

References

1. *Language and Technology*, ESCC-EC-EAEC, Brussels, Luxembourg, 1996.
2. Roukens, J.(1998). The multilingual information society. *Elsnews. The Newsletter of the European Network in Language and Speech*, **7**(1).
3. *Survey of Language Engineering Organizations in Central and Eastern Europe*, EC, 1994.
4. Department of General linguistics, Vilnius University, `http://www.flf.vu.lt/bkk-e.htm`
5. Lithuanian Linguistcs Department, Vilnius Pedagogical University, `http://www.vpu.lt/vpu/faculties/lithuanian.htm`
6. Center of computational linguistics, Vytautas Magnus University, `http://donelaitis.vdu.lt`
7. Department of Language History and Dialectology, Institute of Lithuanian Language, `http://www.mch.mii.lt/more/LKI/L_SID.htm`
8. Department of Dictionaries, Institute of Lithuanian Language, `http://www.mch.mii.lt/more/LKI/L_sZ.htm`
9. Institute of Lithuanian Literature and Folklore, `http://neris.mii.lt/research/research.html`
10. Laboratory for Computer Applications in Research, Institute of Mathematics and Informatics, `http://www.elsnet.org/publications/oldsurvey/ LaboratoryforComputerApplicationinResearch LT2600VilniusLithuania/`
11. Philological Department, Siauliai Pedagogical University `http://www.elsnet.org/publications/oldsurvey/Philological Department5400SiauliaiLithuania/`
12. Regcognition Processes Department, Institute of Mathematics and Informatics, `http://neris.mii.lt/mii/mii_engl/skyr_an/aps.htm`
13. Computer Software Department, Kaunas University of Technology, `http://www.elsnet.org/publications/oldsurvey/ ComputerSoftwareDepartmentLT3031KaunasLithuania/`
14. Lithuanian Dialects, Project of Unesco. Chair in Informatics for the Humanities, `http://www.mch.mii.lt/Tarmes/pradzia.htm`
15. Publishing House TEV, `http://www.elsnet.org/publications/oldsurvey/ PublishingHouseTEV2600VilniusLithuania/`
16. Unesco Chair in Informatics for the Humanities, Institute of Mathematics and Informatics, `http://www.mch.mii.lt/UNESCO_CHAIR/Unesco_chair.stm`
17. Fotonija Ltd, `http://www.elsnet.org/publications/oldssurvey/ FotonijaLtd2600VilniusLithuania/`
18. Sekasoft Ltd, `http://www.elsnet.org/publications/oldsurvey/ SekasoftLtd3000KaunasLithuania/`
19. A. Valiulis. Lithuanization of WWW Pages, `http://daugenis.mii.lt/w3Lithuanization/lietuvizacija.htm`

20. A. Valiulis. If your Netscape Navigator do not show Lithuanian fonts, `http://www.is.lt/del_fontu/font1.html#win95`
21. Speech Research Laboratory, Kaunas University of Technology, `http://www.elsnet.org/publications/survey/SpeechResearchLaboratory3042KaunasLithuania/`
22. Department of Phonoscopic examination, Institute of Forensic examination, `http://www.elsnet.org/publications/survey/DepartmentofPhonoscopicLaboratory LT2600VilniusLithuania/`

A. Lipeika is a Doctor of Technical Sciences, a senior researcher of the Recognition Processes Department at the Institute of Mathematics and Informatics and an associate professor at the Radioelectronics Department of Vilnius Technical University. Scientific interests include: processing and recognition of random processes, detection of changes in the properties of random processes, signal processing and speaker recognition.

J. Lipeikienė is a Doctor of Technical Sciences, a senior researcher of the Recognition Processes Department at the Institute of Mathematics and Informatics and an associate professor at the Informatics Department of Vilnius Pedagogical University. Scientific interests include: processing of random signals, including speech signals, robust methods for determination of change-points in the properties of random processes, data compression.

Kalbos inžinerija Lietuvoje

Joana LIPEIKIENĖ, Antanas LIPEIKA

Kalbos inžinerija, susidedanti iš kalbos ir šnekos apdorojimo, yra labai svarbi kiekvienos tautos daugiakalbėje ir daugiakultūrėje Europoje vystymuisi. Norint nugalėti kalbos barjerus ir naudoti visas kalbas įvairiose žmonių bendaravimo srityse, būtinos priemonės ir sistemos, sukurtos visų tautų kalboms. Šiame darbe aptariama kalbos inžinerijos būklė Lietuvoje.