LITHUANIAN LANGUAGE PROCESSING USING DIGITAL TECHNOLOGIES

Jurgita Mikelionienė
Kaunas University of Technology

Abstract
As Lithuania is about to become a full member of EU, it is important for its national language, although ascribed to regional languages, to stay lively and capable of serving domestic as well as legal, scientific, educational and other needs of Lithuanian speaking society. Having the aim to achieve this, the problem of language computerization is especially relevant now. Digital technologies of Lithuanian language processing – voice recognition, speech synthesis, electronic language data, terminology banks, corpora, speech corpora, computer lexicography, computer software lithuanization and others – are designed and elaborated in the scientific institutes, the institutions of higher education, and private companies; however, there are many unaccomplished works (for example, automated translation), which have to be accelerated and improved. Language is changing very rapidly therefore it is important to record its development and innovations.
KEY WORDS: Lithuanian language, natural language processing, language and speech technologies.

Anotacija
Lietuvai tapus visateise ES nare, svarbu, kad ir jos valstybinė kalba, nors priskiriama regioninėms kalboms, išliktų gyvybinga, pajėgi tarnauti lietuviškai kalbančios visuomenės ne tik buitinėms, bet ir teisinėms, mokslinėms, mokomosioms ir kitoms reikmėms. Taigi kalbos kompiuterizavimo problema dabar yra ypač aktuali. Skaitmeninės lietuvių kalbos apdorojimo technologijos (balso atpažinimo, šnekos sintezės, elektroninių kalbos duomenų, terminų bankų, tekstynų, garsynų, kompiuterinės leksikografijos, programinės kompiuterinės įrangos lituanizavimo ir kitos) kuriamos ir tobulinamos moksliniuose institutuose, aukštosiose mokyklose, privačiose įstaigose, tačiau dar yra daug nenuveiktų darbų (pvz., automatizuotas vertimas), kuriuos reikia spartinti, o tai, kas padaryta, tobulinti. Kalba kinta labai sparčiai, todėl svarbu fiksuoti jos raidą ir naujoves.
PAGRINDINIAI ŽODŽIAI: lietuvių kalba, natūralios kalbos apdorojimas, kalbos technologijos.

Introduction

Many Lithuanian people see almost all problems related to entering the EU more relevant than language cherishing and preservation. The language problem was probably more escalated before the referendum on joining the EU: eurooptimists were claiming that there is no threat for our language; eurorealists or skeptics witnessing almost unrestricted flow of English doubted and feared for the loss of national identity. The same like having received the gifts of socialism one favors the ideas of free market, after general strained russification one sometimes admires the English language too unadvisedly.

There exists an uncertainty about the real future, since the history does not know a case of one political-economic unit comprising 25 independent states with twenty languages of equal status. These numbers force to think not only the citizens and officers of each country separately but EU lawmakers announcing decisions related to language policy as well.

The aim of this article is to define present EU language policy briefly, to reveal a problematic situation of "minor" languages, to reflect upon their opportunity to be competent partners of the English language. Language computerization (its means is the main object of the analysis) is perceived as one of the most realistic ways to preserve a mother tongue; consequently, one of the main objectives of the article is to show the possibilities and tendencies to process the Lithuanian language using digital technologies, the possibilities that are used by the majority of regional as well as international languages.

The necessity to computerize language is dealt with in conferences (On April 21–22, 2004 the conference "Human Language Technologies. The Baltic perspective", which is organized every ten years, was held in Riga, Latvia. On Juni 18–20, 2004 the conference "Languages, technologies and culture diversity" was organized in Kaunas University of Technology), seminars (COLING[1]), scientific articles as well as in articles meant for general public (Danielsson, Utka, 2003; Zinkevičius 2000; Dagienė, 1998), the oficial documents of the European Union (Council Decision of 22 December 2000 adopting a multiannual European Community programme to stimulate the development and use of European digital content on the global networks and to promote linguistic diversity in the information society[2]) and even in the programs of political parties. However, this article is likely to be the first attempt to survey the development of the most important language technologies in Lithuania, arising problems and even threats, in case the works of language engineering are not performed in time.

This article also deals with the branches of Lithuanian linguistics where language technologies have already borne some concrete fruit and with those that lack such kind of products a lot.

1. The Lithuanian Language – Competent Partner of EU Languages

The basis of EU language policy is declared equality of all languages from the perspective of culture variety. This proposition has been exercised since 1958 when the languages of all six Economic community members – there were 4 at the time: French, German, Italian, and Dutch – were announced to be official and working ones. Later, when nine more states had joined the EU, the list of EU official languages lengthened to 11. It was complemented with English, Danish, Greek, Spanish, Portuguese, Swedish and Finnish languages. Since

[1] International Committee on Computational Linguistics organises the International Conference on Computational Linguistics.
[2] *Official Journal,* L 014, 18/01/2001, p. 0032–0040.

Strasbourg Charter in 1992 an exceptional attention to less frequently used languages has been declared: the right to use them in public and privately was proclaimed; *Universal declaration of linguistic rights* in 1996 (Barcelona) started protecting even languages of migrants if they formed any linguistic community. Thus it concerns the rights of all languages in one territory. Unfortunately, one can only envy official languages attention and support, especially if they are used less frequently. We cannot claim strongly that the Lithuanian language will be stranded after its joining the European Union, it seems, though, that the matter of language survival will only concern individual states but not the Union. National self-consciousness alone is unlikely to help here. To say the truth, a lifebuoy is thrown: the most modern, yet the most difficult to implement, right is reminded to have the technologies that would allow to use the information technologies in own languages. In computer science, even "major" languages use universally employed English to save their languages from downgrading. This kind of technologies would safeguard self-expression, education, translation, and other spheres, where official language is used. Therefore if we want the Lithuanian language to be equal partner with other EU languages, to have equal official, working status, the problem of natural language processing with digital technologies is becoming very relevant. Unless it is solved in time, the issue of language equality may become very sore, and some languages may become more equal than others. Not having exercised the right to have information technologies of our own language, though seeking to keep with public advance, we will have to use those of foreign languages in this way helping to annihilate our language.

What is the situation with digital technologies in the Lithuanian language, when Lithuania is about to pass the threshold of the European Union? Isn't Lithuanian implemented too slowly into electronic media? What is state support and attention to language computerization? These questions are topical at the moment; we will try to analyze them briefly.

2. Necessity and reality of Lithuanian language computerization

It should be noted that Lithuania is one of few world countries having its Law of Official Language (VLKK, 1998); *Project of guideline means for Lithuanian language policy* has been ratified recently (VLKK, 2004). To solve the problems analyzed in the article, the following ones are especially relevant: prepare amendments to legal acts, which will set the requirements for the usage and grammaticality of the Lithuanian language in a public electronic media; formulate and subject the program of the Lithuanian language in the information society (works of Lithuanian language analysis, means of automated translation, lithuanization of software, the link between people and computers, etc.) for approval of the government of the Republic of Lithuania; organize international seminars on the issues of functioning of little used EU official languages; design a specialized programming tools to digitalize Lithuanian data bases and archives, etc. (VLKK, 2004). Therefore focus on the official language in Lithuania is declared; moreover, the government has decided to support the programs that solve particular linguistic problems financially. This should be related not only to our living in the age of information and digital technologies, but also to real insights and experience gained from old members of the European Union, where language processing with information technologies is perceived as the only way to preserve a national language (even if it is official one). Items without technological support usually remain only in spoken language, thus in a short-term usage. In this case we will have nothing to pass from generation to generation.

A language is changing rapidly. It is difficult to predict its further developments, yet the prognoses related to language authenticity and purity are not joyful: new words flow into the language, old ones acquire new meanings, a sentence structure is changing – the world round us is changing; our thinking is changing together with the language. Therefore it is important for digital technologies to be able to accumulate, store, and analyze the data of natural language, record changes and novelties. In case of danger for language extinction, such technologies would serve as a pulmotor sustaining the life of the language. Consequently, we should constantly improve and complement our resources and programs, capable of reflecting language development and variety.

Mainly the scientist from Institutes of Computer Science and Mathematics, Lithuanian language, Vilnius University, Vytautas Magnus University, and Kaunas University of Technology design language technologies in Lithuania. Some private companies („Fotonija", „Tilde") are working in this sphere as well. Lithuanian computerization principally is coordinated by the State Commission of the Lithuanian Language as well as Information Society Development Committee.

Much has been done in the last decade and even more unrealized ideas will be materialized when implementing the program „The Lithuanian Language in Information Society 2000–2006" passed by the Government. The main aim of this program – to make sure that, having the aim to preserve the Lithuanian language, it is actively used together with other EU states' languages in the process of integration into the European Union; objectives: the translation of EU documents into Lithuanian and vice versa, speech recognition and synthesis, computer software lithuanization, preparation of linguistic resources, etc.

3. Lithuanian language technologies

- One of the spheres where much work is going on is the installation of Lithuanian primitives into computer systems. This activity comprises and will comprise the creation and coding of stressed letters and other writing signs, the installation of a new Lithuanian keyboard (Tumasonis, Grigas, 2000), the transfer of Lithuanian peculiarities in

information technologies, the localization of open text software, etc. Since the development of original Lithuanian software would not be possible to combine with usual international cooperation, besides it would be complicated financially[3], the Lithuanianization of programs is very important when striving to make a computer suitable to use in Lithuanian cultural and linguistic environment. The dialogue presented in a consumer's language should be a natural phenomenon. The Lithuanian language is still not more peculiar than other ones in the field of program localization (alphabet, order of letters, punctuation or writing of other ideographic signs); and the European medium in general purpose program localization has not been reached yet.

It is symbolic, that a new, original script „Palemonas", corresponding to particularity of Lithuanian writing, appears in the year of Language and book, when celebrating the 100th anniversary of Lithuanian writing retrieval. General standard codes are created for dialects and old writings as well as for phonetic transcription. When creating a font adapted to the system of the Lithuanian language, it was referred to the Renaissance Latin font. The name Palemonas is a symbolic link between the theory of Lithuanian origin, which was popular in the 16th century, and the new original Lithuanian font.

- Application of voice technologies, as the most natural way of communication, to the Lithuanian language should be mentioned separately. Perfect speech recognition, adjusted to dictation and other systems, is pursued. The development of two spheres of voice technologies – automatic identification of speech units and text reading vocally (speech synthesis) – particularly depends on good knowledge in the Lithuanian language. It would be naïve to hope that someone outside Lithuania could create such technologies successfully. Partially, it depends on certain unique accentual, intonational, and grammatical features of the Lithuanian language. Voice technologies are a composite part of the European linguistic infrastructure (Danzin, 1992) and therefore their development is extremely important.

We already have a mediocre synthesizer of Lithuanian, still there is much to edify, especially intonation. The synthesizer of the Lithuanian language „Aistis" was developed at Vilnius University in 1996 (Kasparaitis, 2001). It, as well as an improved variant „Aistis-2" (2003), was primarily intended for computer users with sight disability. The synthesizers read the textual information that appears on the screen vocally. The synthesizers use the segments of natural announcer's speech. There are the syllabic and accentual algorithms of Lithuanian words (Kasparaitis, 2000, 2001a) and the rules for text transcription installed into them (Kasparaitis, 1999). Their word comprehensibility is almost 90 per cent.

Voice technologies are used not only to help to the disabled, but also, to improve the quality of mother-tongue or foreign language learning, for needs of criminalists (Lipeika, Lipeikienė, Telksnys, 2002) specialists of telecommunication (Rudžionis, Ratkevičius et al., 2003), culture heritage and other spheres. Voice technologies contribute to management and protection of linguistic values.

- In order to train people to work with voice recognition systems successfully, it is necessary not only international experience in this sphere, not only powerful Lithuanian factors, like syntax and phonetics, but also to have huge authentic resources – speech corpora (Rudžionis, Žvinys, 2001). There is a lack of comprehensive speech corpora, i.e. bases of natural spontaneous spoken language. It is quite problematic to accumulate as well as to annotate them, although the latter is based on the globally acknowledged Hidden Markov Models or Artificial Neural Network models (Deller, Proakis, Hansen, 2000; Raškinis, Raškinienė, 2003). The first Lithuanian signal data basis is LTDIGITS speech corpus (Rudžionis, Rudžionis, Žvinys, 2000), which, to say the truth, is not characterized by a wide variety of phonetic units and words. In 2002, In the Faculty of Computer Science at Vytautas Magnus University, a universal, annotated colloquial speech corpus was created out of 731 words spoken in isolation (Raškinis, Raškinis, Kazlauskienė, 2003) however, it does not equal the analogous language technologies that other members of the European Union have.
- When recording the situation of a modern language, a huge corpus of the Lithuanian language – a collection of processed electronic texts, with more than 100 mln words at present – is being accumulated at the Vytautas Magnus University[4]. The corpora of written language are necessary for general linguistics as well as for special research in computer lexicography, terminology, even in cultural or social fields (Teubert, 1996; Sinclair, 1999; Biber, 1993; Ido, 1994; Atkins, Fillmore, Johnson, 2003). To satisfy the needs of translation, bilingual English-Lithuanian corpora were started to create (Marcinkevičienė, 2000a, 2000b; Utka, 2004). Composing of comparative corpus that is of new type in the Lithuanian linguistics has been started at Kaunas University of Technology. This is a bilingual Lithuanian-English corpus that reflects the usage of one functional style – modern scientific language. Thematics of the texts is concrete and strictly limited – language of technology (chemistry, logistics, electricity, informatics, etc.) sciences (Mikelionienė, 2002).

3 www.likit.lt

4 http://donelaitis.vdu.lt

- A separate group of language technologies comprises those that are needed to linguists who wish to collect, organize and analyze language facts quickly and credibly, whether they were writing monuments or even the newest elements of lexicon (Mikelionienė, 2000). The demand for such products, the possibility to create new kinds of dictionaries and grammar books, to verify certain hypotheses conditioned the change of some classical branches of linguistics – morphology, syntax, lexicography, etc. – into applied ones as computer morphology and the like. On the other hand, the analyzers of various language levels are necessary to all the systems performing the functions of natural linguistic data search and analysis due to their possibility to be applied.

To perform research in Lithuanian grammar, *Lemuoklis* was created. *Lemuoklis* automatically describes Lithuanian written word forms in grammatical (morphological) aspect and identifies title forms to words, which are called lemmas. The weakness of the process is the ambiguity of lemma creating. When analyzing the typology of morphological polysemy it was determined that up to 40 percent of forms identified automatically in the morphologically annotated corpus are polysemous (Rimkutė, 2003: 76). The author of the morphological analyzer assumes that the problem might be reduced if the regularities of syntactic links among words were formalized (Zinkevičius, 2000, p. 262).

Lithuanian lexicography is becoming more and more computerized as well. Its dependence on computer technologies is getting more and more obvious. The main advantage of computer dictionaries is the possibility to update their data basis and in this way to record the development of the lexis. The most significant work in linguistic aspect is a public electronic version of the Contemporary Lithuanian Dictionary[5]. Using this dictionary one may check the spelling, accentuation and meaning of words and other things important not only to linguists but also to all the users of the language. Nowadays people use bilingual translating dictionaries "Led" and especially "Alkonas" more widely, the latter being an electronic version of "The Great English-Lithuanian" dictionary, which is also an improved version, due to the possibility to perform bidirectional search of a word wanted. Recently, a compact disc with a new international word dictionary "Interleksis" has appeared[6]. Several term dictionaries of certain spheres such as social security[7], computer science[8], hydrogeology, NATO, etc. can be found in the websites of the institutions interested in the field. Lithuanian term basis – term corpus[9] – should serve the needs of term usage and management.

- We are far behind in the field of computerized translation. We are going to feel this lack very soon. Even now in Brussels people talk that document translation into 11 languages is a great burden; and if the translation into 20 languages is needed, the burden may become unbearable. Disorder in institutionary work and enormous expenditure is predicted. That may be the reason why an independent (although initiated by the EU) group "Europa Diversa" (Diverse Europe), founded in 2000 in Barcelona, prepared the project of Linguistic Proposals to European Future that is presented to the convent of future Europe. Those proposals are based on the attitude that, although following the principles of EU organization, it is necessary to facilitate institutional work and consequently not to give official and working status to all languages. If this project is approved, the strengthening of position and preservation of a national language will be the matter of a state itself – it is a painful truth that saving of the drowning is the matter of the drowning. Countries, the languages of which are attributed to less frequently used ones, understand this and work with their shirt sleeves tucked up in the sphere of automated translation: Bulgarians and Czechs have already created computer translators, Hungarians are trying to adapt SYSTRAN system. Sadly, Lithuanians are just making their first steps (Tamulynas, 2003). So that they were successful, we need morphology, syntax and semantics, created on the grounds of corpora. Otherwise there will not be automated distinction of word meanings, semantic annotation, and other things related to automated translation.

Conclusion and proposals

Since the motivation of computer language technologies is clear enough, we have specialists (not enough of them though) and we will have more of them in the future (two-level university studies of computer linguistics have already been introduced), we only have to start working. The activity should develop in two directions:

1) create, organize and sustain computer language resources using not only corpora, but electronic archives that are constantly updated as well;

2) create electronic language standards, which a single company or university is not capable of – we need a special commission.

The attitude of society to computer science lithuanization is ambiguous. We often hear reproaches about quasi-unnecessary, burdensome activity that changes conventional things. Nevertheless we have to realize that, having entered the European Union, we will enhance European identity and not our national one if the most important works of language adaptation to digital technology are not implemented.

The situation is likely to become similar to communicating vessels: if we do not fill our own vessel, others will not do this instead of us, not in Lithuanian. Maybe we should start from the promotion of the Lithuanian

[5] http://www.autoinfa.lt/webdic/
[6] www.fotonija.lt
[7] http://sec.lt/pages/zodynas/1.htm
[8] http://aldona.mii.lt/pms/terminai/term/
[9] http://www.terminynas.lt

language – our language is the most archaic of all living Indo-European languages.

References

Atkins, S. (2003). Lexicographic Relevance: Selecting Information From Corpus Evidence. *International journal of lexicography 16 (3)*: 251–280. Oxford University Press.

Biber, D. (1993). Using register-diversified corpora for general language studies. *Computational linguistics 19(2)*: 219–241. Cambridge: MIT Press.

Dagienė, V. (1998). Šiuolaikinės informacinės technologijos švietime: kalbos problema. *Lituanistika pasaulyje šiandien: darbai ir problemos 3*: 55–64. Vilnius: Baltos lankos.

Danielsson, P., Utka, A. (2003). Academic research and Standards: a Discussion on Standards for Multilingual Language Resources. *Corpus Linguistics,* p. 35–42. Lancaster University.

Danzin, A. (1992). *Towards a European Language infrastucture Strategic Planning Study Group for the Commission of the European Communities* (DG XIII), 31 March 1992. Luxembourg.

Deller, J. R., Proakis, J. G. (1999). *Discrete-time processing of speech signals*. New York: IEEE Press classic reissue.

Ido, D., Alon, I. (1994). Word sense disambiguation using a second language monolingual corpus. *Computational linguistics 20 (4)*: 563–596. Cambridge: MIT Press.

Kasparaitis, P. (1999). Transcribing of the Lithuanian text using formal rules. *Informatica 10(4)*: 367–376. Vilnius: MII.

Kasparaitis, P. (2000). Automatic stressing of the Lithuanian text on the basis of a dictionary. *Informatica 11(1)*: 19–40. Vilnius: MII.

Kasparaitis, P. (2001). Computer Synthesizer of Lithuanian Language "Aistis". *Sound Blaster 2001* (CD). Kaunas: KTU leidykla.

Kasparaitis, P. (2001a). Automatic stressing of the Lithuanian nouns and adjectives on the basis of rules. *Informatica 12(2)*: 315–336. Vilnius: MII.

Lipeika, A., Lipeikienė, J., Telksnys L. (2002). Development of Isolated Word Speech Recognition System. *Informatica 13(1)*: 37–46. Vilnius: MII.

Marcinkevičienė, R. (2000a). Tekstynų lingvistika. Teorija ir praktika. *Darbai ir dienos 24*: 7–64. Kaunas: VDU leidykla.

Marcinkevičienė, R. (2000b). Patterns of words usage viewed by corpus linguistics. *Kalbotyra 49(3)*: 71–80. Vilnius: VU leidykla.

Mikelionienė, J. (2000). Šiuolaikiniai metodai kalbos naujovėms tirti. *Darbai ir dienos 24*: 65–73. Kaunas: VDU leidykla.

Mikelionienė, J. (2002). Palyginamojo tekstyno kūrimo principai, problemos ir panaudojimo galimybės. *Kalbų studijos 3*: 55–59. Kaunas: Technologija.

Raškinis, A., Raškinis, G., Kazlauskienė, A. (2003). VDU bendrinės lietuvių šnekos universalus anotuotas garsynas. *Informacinės technologijos 2003*: konferencijos pranešimų medžiaga. Kaunas: Technologija, p. IX, 28–34.

Raškinis, G., Raškinienė, D. (2003). Development of Medium-Vocabulary Isolated-Word Lithuanian HMM Speech Recognition System. *Informatica 14(1)*: 75–84.

Rimkutė, E. (2003). The typology of morphological ambiguity. *Lituanistica 4(56):* 60–78. Vilnius: LMA.

Rudžionis, A., Žvinys P., Rudžionis, V. (2001). Lietuvių kalbos garsynas: projektavimas, patirtis ir tolesnės plėtros perspektyvos. *Informacijos mokslai 17:* 77–84. Vilnius: VU leidykla.

Rudžionis, A., Rudžionis, V., Žvinys, P. (2000). Lietuvių šnekamosios kalbos garsynas LTDIGITS: rezultatai ir problemos. *Informacinės technologijos 2000*. Kaunas: Technologija, p. 162–166.

Rudžionis, A., Ratkevičius, K., Rudžionis, V., Kasparaitis, P. (2003). Voice operated informative telecom services. *Elektronika ir elektrotechnika 3(45)*: 17–22. Kaunas: Technologija.

Sinclair, J. (1999). The Computer, the Corpus and the Theory of Language. *Lingua 1*: 24–32.

Tamulynas, B. (2003). Multilingual computer-based communication and language processing: Lithuanian case. *Computational linguistics and intellectual technologies*. International conference proceedings. Moscow (Protvino, June 11–16), p. 663–666.

Teubert, W. (1996). Comparable or parallel corpora. *International journal of lexicography 9:* 218–237. Oxford University Press.

Tumasonis, V., Grigas, G. (2000). Naujos lietuviškos kompiuterio klaviatūros standartas. *Informacijos mokslai 14*: 105–112. Vilnius: VU leidykla.

Utka, A. (2004). Phases of Translation Corpus: Compiliation and Analysis. *International Journal of Corpus Linguistics 9–2*. Amsterdam: John Benjamins forthcoming.

VLKK (1998). Lietuvių kalbos komisijos nutarimai 1977–1998. *Valstybinė lietuvių kalbos komisija prie Lietuvos Respublikos Seimo*. Vilnius: MELI.

VLKK (2004). *Valstybinės kalbos politikos gairės (priemonių planas)*. Valstybinė lietuvių kalbos komisija. Prieiga Internete: www.vlkk.lt

Zinkevičius, V. (2000). Lemuoklis – morfologinei analizei. *Darbai ir dienos 24:* 245–273. Kaunas: VDU leidykla.

Gauta 2005 02 24

Pasirašyta spaudai 2005 03 07

Spausdinti rekomendavo: doc. dr. A. Kazlauskienė, prof. K. Kriščiūnas

LIETUVIŲ KALBOS APDOROJIMAS SKAITMENINĖMIS TECHNOLOGIJOMIS

Jurgita Mikelionienė

S a n t r a u k a

Daugeliui Lietuvos gyventojų su buvimu Europos Sąjungoje susijusios problemos atrodo daug aktualesnės už kalbos puoselėjimą ar saugojimą. Apie Europos Sąjungos nuostatą skatinti nacionalinių kalbų vartojimą šiuo metu kalbama tikrai mažiau nei apie kitas, kurios susijusios su naujų šalių narių žemės ūkio sektoriaus ar darbo rinkos atvėrimo problemomis. Lietuvai tapus visateise ES nare, svarbu, kad ir jos valstybinė kalba, nors ir priskiriama regioninėms kalboms, išliktų gyvybinga ir pajėgi tarnauti lietuviškai kalbančios visuomenės ne tik buitinėms, bet ir teisinėms, mokslinėms, mokomosioms ir kitoms reikmėms. Šiuo atveju ypač aktualus kalbos kompiuterizavimas.

Straipsnyje glaustai aptariama šiandienė ES kalbų politika, atskleidžiama problemiška „mažųjų" kalbų situacija, svarstoma jų galimybė būti visateisėmis anglų kalbos partnerėmis. Kalbos kompiuterizavimas (jo produktai yra pagrindinis šio straipsnio objektas) suvokiamas kaip vienas realiausių gimtosios kalbos išsaugojimo būdų, todėl vienas straipsnio uždavinių – parodyti lietuvių kalbos apdorojimo skaitmeninėmis technologijomis galimybes ir tendencijas, kuriomis naudojasi daugelis ne tik regioninių, bet ir tarptautinių kalbų.

Apie kalbos kompiuterizavimo būtinybę kalbama konferencijose, seminaruose, moksliniuose ar eiliniam skaitytojui skirtuose straipsniuose, oficialiuose Europos Sąjungos dokumentuose, net politinių partijų programose. Bet šis straipsnis – bene pirmasis mėginimas apžvelgti svarbiausių kalbos technologijų kūrimo situaciją Lietuvoje, aptarti iškylančias problemas ir net grėsmes, jei laiku nebus atlikti kalbos inžinerijos darbai.

ES kalbos politikos pagrindas – deklaruojamoji visų kalbų lygybė žvelgiant kultūrų įvairovės aspektu. Šis teiginys pradėtas įgyvendinti nuo 1958 metų, kai visų 6 Ekonominės bendrijos narių kalbos (jos tuo metu buvo 4: prancūzų, vokiečių, italų, olandų) paskelbtos oficialiomis ir darbinėmis. Nuo 1992 m. „Strasbūro chartija" dekla-

ruoja, kad būtina daugiau dėmesio skirti rečiau vartojamoms kalboms: skelbiama teisė jas vartoti viešai ir privačiai, o 1996 metų Barselonos „Visuotine kalbinių teisių deklaracija“ pradedama proteguoti net migrantų, jei tik jie sudaro kalbinę bendruomenę, kalbas. Taigi kalbama apie visų kalbų, vartojamų vienoje teritorijoje, teises.

Norint, kad lietuvių kalba būtų lygiateisė kitų Europos Sąjungos kalbų partnerė, turėtų vienodą oficialų, darbinį statusą, aktuali tampa natūralios kalbos apdorojimo skaitmeninėmis technologijomis problema. Jei nepasinaudosime teise turėti savas kalbos informacines technologijas, siekdami neatsilikti nuo visuomenės pažangos, turėsime naudotis svetimkalbėmis ir taip patys pradėsime naikinti savo kalbą. Į kalbą plūsta nauji žodžiai, praplečiamos senųjų reikšmės, keičiasi sakinio sandara – mainosi pasaulis, kinta mąstymas, kartu ir kalba. Todėl svarbu, kad skaitmeninės technologijos būtų pajėgios kaupti, saugoti, nagrinėti natūralios kalbos duomenis, fiksuoti permainas ir naujoves. Patirtis, perimta iš Europos Sąjungos šalių senbuvių, kuriose kalbos apdorojimas informacinėmis technologijomis suprantamas kaip vienintelis nacionalinės, net jeigu ji ir valstybinė, kalbos išsaugojimas būdas, yra naudinga. Lietuvių kalbos politikos kūrėjai puikiai suvokia, kad technologijų nepalaikomi dalykai liks tik šnekamojoje kalboje, taigi jų vartosena bus neilgalaikė.

Kalbos technologijas Lietuvoje daugiausia kuria Informatikos ir matematikos, Lietuvių kalbos institutų, Vilniaus, Vytauto Didžiojo ir Kauno technologijos universitetų mokslininkai. Šioje srityje darbuojasi ir kelios privačios įstaigos, pvz., „Fotonija“, „Tildė“. Viena sričių, kur intensyviai darbuojamasi – lietuvių kalbos bazinių elementų diegimas kompiuterinėse sistemose. Ši veikla apima ar ateityje apims lietuviškų kirčiuotų raidžių ir kitų rašto ženklų aibės sudarymą, kodavimą, naujos lietuviškos klaviatūros įdiegimą, lietuvių kalbos ypatybių perkėlimą į informacines technologijas, atvirojo teksto programinės įrangos lokalizavimą ir kt. Kol kas, nors lietuvių kalba lokalizuojant programas (raidyną, raidžių rikuotę, punktuaciją ar kitų ideografinių ženklų rašybą) ir neišsiskiria iš kitų, europinis bendros paskirties programų lokalizacijos vidurkis dar nepasiektas. Vienas naujausių darbų, gražiai sutapusių su lietuvių kalbos rašto atgavimo 100 metų jubiliejumi, – lietuviško šrifto „Palemonas“ sukūrimas: tarmių, senųjų raštų rašmenims, fonetinei transkripcijai sukurti bendri standartiniai kodai.

Minėtinas balso technologijų, kaip natūraliausio bendravimo būdo, taikymas lietuvių kalbai. Siekiama tobulo šnekos atpažinimo, pritaikyto diktavimo ir kt. sistemose. Balso technologijos diegiamos ne tik siekiant pagerinti gimtosios ar užsienio kalbos mokymosi kokybę, jos taip pat padeda neįgaliesiems, teisėsaugininkams, telekomunikacijų, kultūros paveldo ir kt. sričių specialistams. Lietuvių kalbos sintezatorius „Aistis“ jau sukurtas, reikia tik jį patobulinti, ypač intonacijos sritį. Naivu būtų tikėtis, kad kas nors ne Lietuvoje galėtų sėkmingai kurti balso technologijas. Tai iš dalies susiję su tam tikromis unikaliomis lietuvių kalbos kirčiavimo, intonacinėmis ir gramatinėmis lietuvių kalbos savybėmis. Šios technologijos yra Europos kalbinės infrastruktūros sudėtinė dalis, todėl jų plėtra yra labai svarbi.

Fiksuojant dabartinės kalbos situaciją, Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centre kaupiamas didžiulis lietuvių kalbos tekstynas – programiškai apdorotas elektroninių tekstų rinkinys, šiuo metu jau viršijęs 100 mln. žodžių. Rašomosios kalbos tekstynai būtini ne tik bendrosios kalbotyros, bet ir specialiesiems, pvz., kompiuterinės leksikografijos, terminologijos, net kultūrologiniams, sociologiniams tyrimams. Vertimo reikmėms pradėti kurti dvikalbiai tekstynai: KTU renkamas technologijos mokslų srities lietuvių ir anglų kalbų panašios tematikos tekstų tekstynas, leisiantis nagrinėti, lyginti elektrotechnikos, informatikos, mechanikos ir kt. inžinerijos sričių terminus, jų vertimo atitikmenis, kurti terminų žodynėlius. Labai trūksta garsynų, t. y. natūralios, spontaniškos sakytinės kalbos bankų. Tokie išsamūs šnekamosios kalbos tekstynai leistų sėkmingai kurti kompiuterio valdymo komandų atpažinimo sistemas.

Atskirą grupę sudaro tokios kalbos technologijos, kurios pirmiausia būtinos lingvistams, norintiems patikimai ir greitai rinkti, sisteminti, nagrinėti kalbos faktus, ar tai būtų raštijos paminklai, ar net patys naujausi žodyno elementai. Gramatikos reikmėms tenkinti sukurtas morfologijos analizatorius „Lemuoklis“. Galima pasidžiaugti keliolika elektroninių žodynų.

Lietuviai dar tik žengia pirmuosius žingsnius kompiuterizuojant vertimą. Kad jie būtų sėkmingi, reikia morfologijos, sintaksės, semantikos, sukurtos remiantis tekstynais. Kitaip nebus žodžių reikšmių automatinio skyrimo, semantinio anotavimo ir kt. su automatizuotu vertimu susijusių dalykų. Briuselyje jau kalbama apie dokumentų vertimo į visas 20 Europos Sąjungos kalbų problemas, kurios gali sutrikdyti institucijų darbą, be to, tai didžiulės išlaidos. „Kalbinių pasiūlymų Europos ateičiai“ projektas (jį parengė Europos Sąjungos iniciatyva susikūrusi grupė „Europa diversa“) remiasi nuostata, kad negalima visoms kalboms suteikti oficialių ir darbinių kalbų statuso. Vienas būdų jį gauti – automatizuoti vertimą.

Kadangi kompiuterinių kalbos technologijų kūrimo motyvacija yra gana aiški, specialistų turime (tiesa, nepakankamai) ir ateityje jų turėtų daugėti (pradėtos universitetinės abiejų pakopų kompiuterinės lingvistikos studijos), belieka imtis darbo. Jis turėtų vykti dviem kryptimis:

1) kompiuterinių kalbos išteklių kūrimo, tvarkymo ir palaikymo ne tik tekstynais, bet ir elektroniniais nuolat atnaujinamais dokumentų archyvais;
2) elektroninės kalbos standartų nustatymo, ko jokia įmonė ar universitetas nepadarys – gal net reikia specialios komisijos.

Visuomenės požiūris į kompiuterijos lietuvinimą nėra vienareikšmis. Tačiau turbūt reikia įsisąmoninti, kad įstoję į Europos Sąjungą europinį identitetą mes sutvirtinsime, tačiau tautinį, jeigu laiku nebus įgyvendinti svarbiausi kalbos pritaikymo skaitmeninėms technologijoms darbai, – vargu. Gal pirmiausia reikėtų pradėti nuo lietuvių kalbos reklamos, juk mūsų kalba – archajiškiausia iš visų gyvųjų indoeuropiečių kalbų.