The Development of Speech Technologies in Serbia Within the European Research

Nataša Vujnović Sedlar*, Slobodan Morača*, Vlado Delić*
* Faculty of Technical Sciences, Novi Sad, Serbia
natasav@uns.ac.rs, moraca@uns.ac.rs, vdelic@uns.ac.rs

***Abstract*—This paper presents potentials, needs and problems of creating an adequate research area for the development of speech technologies in Serbia within the European Research Area. These technologies are a very important part of the strategy Europe 2020, because of Europe's awareness that speech technologies are one of the missing parts that will help Europe to build a unified digital market, overcome language barriers and increase the fluctuation of employees. The authors of the paper made a comparative analysis of the European actions in the European Research Area and projects in Serbia dedicated to the development of speech and language technologies. The analyses shows how European programs influence the development of speech technologies in Serbia.**

***Key words* - European Research Area, European programs, Europe strategy, speech and language technologies**

I. INTRODUCTION

One of the characteristics of the European Union, as a multicultural and multinational community, is the diversity of languages that are spoken within its border. From the beginning, European leaders have been aware of that and presented the European languages as an inalienable part of its cultural heritage. Besides these, there is also a political reason why Europe keeps language diversity as an integral part of its policy since its establishment [1]. However, the linguistic diversity requires a significant investment and large budget funds have been spent on translation services only to ensure the availability of formal documents in 24 [2] official languages. On the other hand, it is of the great importance to ensure access and usability of information in the mother tongue languages to all European citizens.

Language diversity in Europe is supported by various European policies. In the line with the Multilingual Policy [3], Europe supports various activities in education, research programs, language learning and development of language and speech technology. The development and implementation of speech and language technologies, which should break languages barriers between European citizens and push raising European economy, have been helped through the Information Society Policy and the Research and Innovation Policy. Regarding to language and speech technologies, the Policy of Information Society is in correlation with the Multilingual Policy [4], or it can be said that these two policies have the same aim to provide the availability of content in different languages at all European languages and wider across all communication channels and sources of information.

European Commission presume that language and speech technology will bring new quality of social and business life for citizens of Europe and because of that includes development of these technologies in some of the European programmes to ensure market introduction technology readiness level while these technologies should have great influence at business collaboration as well as staff fluctuation, better knowledge sharing etc. [5].

That's why Europe has already invested significant resources in the development of these technologies through the Framework Programme and the entire European Research Area and still does it. The subject of this paper is to make correlations between the projects for development of speech and languages technologies in Serbia and project supported in the European research area from this topic and to understand the impact that the European research area has on development of these technologies in Serbia. Also, this paper indicates the weak points of the Serbian project and suggests improvements for better involvement of the Serbian projects in European research area.

II. LANGUAGE AND SPEECH TECHNOLOGIES

The human language appears in both oral and written form. Speech as an oral form of the language is the oldest and most natural way of verbal communication. On the other hand, text like written form of language allows storage of human knowledge and keeps it from forgetting. Language technology deals with language aspects - sentences, grammar and meaning of sentences in the field of information communication technology are important for the development of speech technologies and word processing. It is a branch of ICT technology that is based on knowledge of linguistics and other interdisciplinary fields of science like mathematics, computer sciences, telecommunications, signal processing, and others.

A. Language, speech and IT technologies

Voice machines have intrigued humans for a long time. According to the current knowledge, the first attempts of creating such machines dating from the 13th century, when the German philosopher Albertus Magnus and the English scientist Roger Bacon [6, 7] made metal heads that produce voice separately from each other. Todays in the era of digitalization, the interest of scientists for the realization of the machines that will understand, recognize speech and translate it from one language to another is much bigger and in relations with the needs of a modern person.

Nowadays, world characterize an abundance of information that is important for different areas of people's lives, not just business life, but also private and social aspects of life. Beside that it should be stressed that information-exchanging channels have been significantly expanded in comparison with the time of first produced voice machines. Information has a global character; there are no limits or borders. Language, as a most natural holder of information, has found itself in a new context, a new environment, which requires study of all aspects of language from the new perspective, the information technology perspective.

Language technology, or as it is often called human language technology (HLT), is the information-communication technology that deals with language, the very complex medium for communication, in a new digital environment. Language technology provides great support to the development of speech technology and text processing technology.

It should also be noted that speech technologies and text processing are intertwining and overlapping with other information communication technologies. For example, multimedia presentation of information collects pictures, music, speech, gestures, facial expression and other forms of information presentations. All that defines the meaning of spoken text and because of that these technologies cannot be studied separately.

In the domain of language technology, researchers are engaged in different research fields such as automatic translation, Automatic Text Summarization, automatic text analysis, optical character recognition, spoken dialogue, speech recognition and speech synthesis. Beside that, researchers have been faced with various problems such as the segmentation of written text, speech segmentation, solving ambiguous meaning of words, syntactic ambiguity solving, overcoming the imperfections of the input data. They also have to take into account the context and the speaker's intentions. One of the biggest problems here is the dependence of these technologies on the language. There is a large degree of inability of applying methods developed for one language to other languages.

B. Development and implementation of language and speech technologies in Europe

The development and implementation of these technologies in Europe differs from language to language. META-NET [8], Network of Excellence dedicated to fostering the technological development of a multilingual European information society, has implemented a series of reports entitled Europe's Languages in the Digital Age [9]. The document treated the state of language and speech technologies for 30 languages, whereby the following areas have been observed:

- Automatic translation - also taking into account: the quality of existing technology, the number of covered language pairs, coverage of linguistic phenomena and domains, the quality and size of the parallel corpus, the amount and variety of applications.
- Speech processing - in which is observed the quality of existing speech technologies, domain coverage, the number and size of the existing corpus, the volume and variety of available applications.
- Text analysis - with an emphasis on the quality and coverage of existing technologies in the field of morphology, syntax and semantics, completeness linguistic phenomena and domains, amount and variety of applications, the quality and size of the corpus, the quality and coverage of lexical resources.
- Resources - the quality and size of text, voice and parallel corpus, the quality/coverage lexical resources and grammar.

Although the language and speech technology can solve the complex issue of multilingualism in Europe, its development is uneven and in some languages such as Lithuanian and Irish it is at the beginning. In addition, it should be noted that the automatic translators are [10] tools that will contribute to the unity of the European market, because it attempts to help bridging the language barrier that currently exist, cannot been fully used yet. Europe's awareness of the necessary development of these technologies allocates significant resources for their development through research funds [11]. A great part of the funds for the development of speech and language technology comes from national funds. These national projects generate very good results that should eventually be implemented and upgraded in European programmes.

The development of language and speech technology not only takes place at research institutions, but also at innovative companies, most of which are small and medium enterprises. There are about 500 European companies that actively participate in the development and/or implementation of language and speech technology. Typically, most of them are focused on national markets and are not included in the European value chain.

III. EU SUPPORT TO THE DEVELOPMENT OF LANGUAGE AND SPEECH TECHNOLOGIES

A. EU support through the Framework Programmes

European Union has thrived to support development of languages technologies and to take leading place in that development through various funding programmes, mostly through the Framework programmes. Within the Framework Programmes it can be identified several major research sub-areas of language and speech technologies that are financed:

- Automated Translation;
- Multilingual Content Authoring & Management;
- Speech Technology & Interactive Services;
- Content Analytics;
- Language Resources;
- Collaborative Platforms.

In order to make overview of the language and speech technologies research directions, supported by European Commission through Framework programmes, authors analysed funded projects available in CORDIS base [13]. Usually analysed projects did not address only one topic, but more similar ones. Relative percentage participation of each topic in relation to total number of funded projects in the last three Framework programmes is presented in the table below.

%	*Automated Translation*	*Multilingual Content Authoring & Management*	*Speech Technology & Interactive Services*	*Content Analytics*	*Language Resources*	*Collaborative Platforms*
FP5 - IST	11,34	41,24	62,89	30,93	32,99	10,31
FP6 - IST	7,41	24,07	79,63	53,70	33,33	5,56
FP7 - ICT	40,00	21,54	29,23	23,08	26,15	15,38

Table 1. Relative percentage participation of each topic

The Automated translation topic, where the projects are mostly devoted to overcoming the machine translation problems and developing different cross-lingual applications, has got up to 40% funding rate in 7th Framework programme, while in 5th and 6th Framework programme funding rate of the approved projects for this topic was 11,34% and 7,41% respectively.

The funding rate for the topic called Speech technology and interactive services has been significantly decreased in last Framework programme. Until the 7th Framework programme it was the most supported topic, but the Automatic translation has taken the lead. In 5th Framework programme this topic was supported in 62,89% projects, in 6th Framework programme in 79,63% projects, while in 7th Framework programme only 29,23% projects with this topic was funded.

Topic Multilingual Content Authoring & Management as a topic of interest for European Union after 5th Framework programme has dropped double of the over all approved projects. Its participation percentage in 5th Framework programme was 41,24% of all funded projects, and then EU support to this topic dropped at 24,07% in the 6th Framework programme. Similar percentage, 21,54%, remained in 7th Framework programme. It is important to stress that in 4th and 5th Framework programme this topic has got special support through specific programmes Multilingual Information Society (MLIS) [14].

Content Analytics as a topic of interest for the development of language and speech technologies, through these three framework programmes, reached its peak in 6th Framework programme, when percentage of funded project related to this topic was 53,70%. Peak in the 6th Framework programme was reached owing to the calls of Semantic-based Knowledge Systems and Proactive Initiatives.

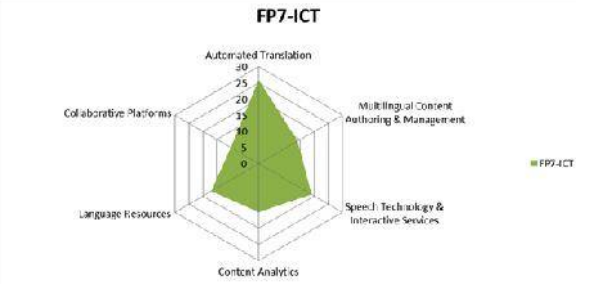

Figure 1. Representation of topics in Framework programme 7

The Language Resources topic has been uniformly supported during each of the three observed framework programs. Nevertheless funding percentages were reached 32,99% in the 5th, 33,33% in the 6th, and up to 26,15% in the 7th Framework programme. Main aim of European Commission by supporting this kind of project was to overcome evident lack of language resources for European languages.

Projects Collaborative Platforms have goal to improve further research to ensure leading position in this domain, or to prepare research community for the next research phase related to the new development directions and the problems linked to multilingualism of Europe and usage of new resources. Percentage partake of this topic in the 5th Framework programme was 10,31%, in the 6th 5,56%, and in the 7th 15,38%.

The analysis represents a big change in direction of support research topics related to language and speech technologies. The primate status of the topic Speech technology and Interactive Services has now been overthrone by the topic Automated Translation. This trend was detected also in the calls established in the Work programme for 2014 and 2015 in new Framework programme Horizon 2020[15]. Enormous change can be seen in the Work programme for 2016 and 2017, where calls explicitly dedicated to the language technologies cannot be registered like in previous calls.

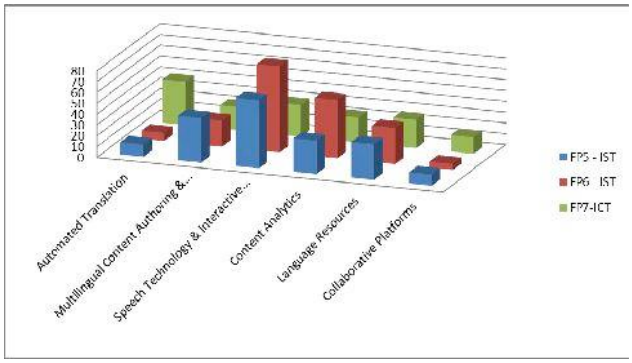

Figure 2. Representation of topics in Framework programmes

IV. The Language and Speech Technologies in Serbia

The development of language and speech technologies in Serbia does not have a long history. Some of the first attempts that deal with computer linguistic were made in 1970's when the first software tool for discovering and correction of spelling mistakes was made. Besides this tool a more recent software RAS has been introduced. The software has processed Serbian text on the computer, which offers correction of the text: dividing words into syllables, conversion of the codes pages and sorting out punctuation.

It should be noticed that the development of language technologies in Serbia is mostly scattered in research organisations. The development is taking place at the faculties of Philosophy as well as at Mathematical and Electrotechnical faculties of the University of Novi Sad and Belgrade. There is a certain level of cooperation between language technology research groups established at these faculties, but with evident problem of correlation among the achieved research foregrounds. The stated cooperations are still insufficient to make a serious progress in this area.

In the area of speech technologies the greatest accomplishment has made by the Faculty of Technical Sciences of the University of Novi Sad. They have managed to put in practice speech synthesis and speech

recognition for Serbian, Macedonian and Croatian languages. Besides that, the Faculty has made accentuation morphological dictionaries for Serbian and Croatian language, for over 4 and 3 million words respectively. The Faculty cooperates closely with AlfaNum company which is making a transfer of these technologies for commercial usage.

V. Language and Speech Technologies for Serbian Language in ERA

The directions of development of language and speech technologies in Serbia match the four topics of interest to the European Union, and to the Speech Technology and Interactive Services, Content Analytics, Language Resources and Collaborative Platforms. The acceptable level of development compared to the technology of English language, as a reference point against which to measure the development of language and speech technologies to other languages, is made only in the area of synthesis. The level of development of tools and resources for Serbian language is quite low, and the volume and quality of text, voice and parallel corpus, and the quality of lexical resources and grammar should seriously be increased. Serbia when it is still working on developing core technologies, and scientific developments outside the domain of relatively rare.

In all of this it should be taken into consideration that the language and speech technology developed in Serbia thanks to national research programmes. Support the development of core technologies through the European Programme occurred in the period when Serbia could legitimately participate in these programs. Scientists from Serbia who work in this specialized field currently take part in collaborative platforms within the Framework Programme, EUREKA program, COST and SCOPES program. Since 2007 there has been a group from the Faculty of Technical Sciences that has submitted seven applications for the participation in the framework programs mainly from the topic of Speech technology and interactive resources, where the two received a score of 13 and 13.5 points, Serbia failed to take significant participation in the ERA.

VI. Discussion

Since 2007, beginning of the Seventh Framework Programme, Europe has changed the priority of topics funded through the Framework Programme. This change of course was retained in the new Framework Programme Horizon 2020. At the moment Serbian researcher have to make an extra effort and try to reorganize the project development of language and speech technologies in order to leverage the research policy in Europe in this field. Also, Serbian researchers have recognized importance of their inclusion and chance for the funding of further research through financial support of the framework programme. In addition they find the exchange of experience and knowledge with researchers from abroad necessary for their further research work.

To realize high quality alignment directions of the development of language and speech technologies Serbia with Europe, it is necessary:

- Increase the amount and quality of text, voice and parallel corpus, and the quality of lexical resources and grammar
- Bring together interdisciplinary teams to improve the situation
- Develop co-ordination of research activities in this field in order to reduce the gap in the development of speech and language technology for Serbian language compared to English, German and French.

Serbia should also seek their chance at commercializing existing technologies whether through direct marketing of advanced technology on the market or through other program of the European Commission funded the research and development of technologies to be quickly commercialized.

Acknowledgment

The presented research was performed as a part of the project “Development of Dialogue Systems for Serbian and Other South Slavic Languages” (TR32035), funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia.

References

[1] Gianni Lazzari (2006), "*Human Language Technologies for Europe*", available at: http://cordis.europa.eu/documents/documentlibrary/90834371EN6.pdf

[2] http://ec.europa.eu/languages/policy/language-policy/official_languages_en.htm

[3] Council of the European Union (2008), "*Council Resolution on a European strategy for multilingualism*", available at: http://www.ceatl.eu/wp-content/uploads/2010/09/EU_Council_multilingualism_en.pdf

[4] European Commission (2010), "*A Digital Agenda for Europe*", available at: http://ec.europa.eu/information_society/digital-agenda/publications/.

[5] The Language Rich Europe Consortium (2011), "*Towards a Language Rich Europe*". Multilingual Essays on Language Policies and Practices, British Council, available at: http://www.language-rich.eu/fileadmin/content/pdf/LRE_FINAL_WEB.pdf.

[6] David Lindsay (1997), "*Talking Head*" Invention. & Technology, pp. 57-63

[7] http://www.haskins.yale.edu/featured/heads/simulacra.html

[8] http://www.meta-net.eu/

[9] http://www.meta-net.eu/whitepapers/overview

[10] LT-Innovate (2013), "*Status and Potential of the European Language Technology Markets*", available at: http://www.lt-innovate.eu/

[11] http://cordis.europa.eu/fp7/ict/language-technologies/

[12] COM(2005) 596 - The 2005 Commission communication A new framework strategy for multilingualism

[13] http://cordis.europa.eu/fp7/ict/language-technologies/portfolio_en.html

[14] Kimmo Rossi (2013), "*Language technology in the European programmes and policies*", 6th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Poznań, Poland, available at: http://www.ltc.amu.edu.pl/pdf/abstract-kimmo-rossi.pdf

[15] http://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/main/h2020-wp1415-leit-ict_en.pdf