Serving Multilingual Europe

Svetla Koeva[1]

[1] Institute for Bulgarian Language "Prof. L. Andreychin", Bulgarian Academy of Sciences
svetla@dcl.bas.bg

1. Introduction

Serving Multilingual Europe is the most precise wording that characterizes the project CESAR (**CE**ntral and **S**outh-East Europe**A**n **R**esources), which was formulated by prof. Tamás Váradi, the coordinator of the project. His paper entitled *Serving Multilingual Europe: The CESAR Project* is the most prominent study included in the book *Language Resources and Technologies for Bulgarian*, published by "Professor Marin Drinov" Publishing House of the Bulgarian Academy of Sciences (Váradi, 2014) in 2014 and dedicated to the results of two major European projects, one of which is the project CESAR. The CESAR project (Central and South-East European Resources)[1] was funded by the European Commission through the ICT Policy Support Programme, Grant agreement no.: 271022. The runtime of the project was from February 1st, 2011 until January 31st, 2013.

At the introduction of his paper, Prof. Váradi presented in depth the rationale motivating the project and its origins, outlining the issue of tackling language as one of the most prominent challenges in the age of digital communication, which is becoming increasingly multimodal and multilingual. Prof. Váradi stressed "the mission of language technologies to ensure that people can communicate with each other as well as with automated services via digital devices in the most natural, unconstrained manner possible" (Váradi, 2014: 9). In fact, the author predicted the extreme need for the collection and processing of big, multimodal and multilingual data, a need that still exists as language technology crucially depends on data, and postulated the requirement for language resources that are applicable, useful and easily available for language technologies. All this motivated the building of a language technology infrastructure in a pan-European coordinated manner, the META-NET (A Network of

[1] http://cesar.nytud.hu

DOI: https://doi.org/10.18135/VT70.8

Excellence forging the Multilingual Europe Technology Alliance: META). The consortium consisted of nine partners covering six languages, with five of them belonging to the Slavic group of languages (Bulgarian, Croatian, Polish, Serbian, Slovak), and the sixth, Hungarian – a Finno-Ugric language. The CESAR project, as part of META-NET, intended to address the issue of the sporadic development of language resources and tools for the so-called less-resourced languages by enhancing, upgrading, standardising and cross-linking a wide variety of language resources and tools and making them available by means of an open infrastructure.

2. META-SHARE and CESAR project

The central objective of the CESAR project was "to produce and make available a comprehensive set of language resources and tools covering Bulgarian, Croatian, Hungarian, Polish, Serbian, and Slovak" (Váradi, 2014: 11). The focus was on diverse types of resources and tools: mono- and multilingual speech databases, mono- and multilingual corpora, dictionaries, wordnets, natural language processing tools: tokenisers, lemmatisers, taggers, parsers, named-entity recognizers and so on. The activities were directed not so much to the creation of new language resources and language technology tools, but rather to "the enhancement of existing resources and tools (in size, coverage, precision, recall, accuracy), the adaptation of resources and tools to become compliant with the agreed standards for interoperability, as well as the upgrade of resources and tools by combining them with other resources and tools" (Váradi, 2014: 12). Here the role of professor Váradi in the preparation of the project's proposal has to be emphasized. Thanks to his skills to coordinate the activities of people with different skills, experience and knowledge; to analyze the objective conditions and to offer the most appropriate solutions in relation to the needs, potentials and expectations; to formulate clearly and precisely the project objectives, work packages, and expected results the successful realization of the project was presupposed and to a great extent predetermined.

Within the CESAR project the META-SHARE platform[2] has been established and populated with language resources and tools (in collaboration with other members of META-NET network of excellence). The META-SHARE platform organises an open network of repositories for

[2] http://www.meta-share.org

sharing language data, tools and web services. As META-SHARE is implemented in the framework of the META-NET Network of Excellence, the CESAR project contributed to the building of META-SHARE and its population with language resources, language processing and annotation tools and technologies, and services. Prof. Váradi explained in his paper that servers linked to META-SHARE form a chain of nodes, organized into a hierarchical structure: managing nodes are synchronized, and provide the META-SHARE metadata and resources, while network nodes are not synchronized, but harvested by a managing node. Within the project, for each language included in the project at least one network node was established and in 2021 the META-SHARE platform itself and many of CESAR network nodes are operating ensuring long-term sustainability (for example, the Hungarian and the Bulgarian nodes). The CESAR project provided through META-SHARE 251 language resources and language processing tools for Bulgarian, Croatian, English, Hungarian, Polish, Serbian, Slovak, and other languages: 66 tools and services; 120 mono- and multilingual corpora, and 65 lexical and conceptual resources. Prof. Váradi stressed within his paper that language resources created, improved and expanded within the Croatian project are at a high technological level, which allows for their direct integration in various applications based on language technologies, both for research and commercial purposes.

To provide the most suitable resources and tools, the project partners have developed a special methodology for the selection of resources. The consortium developed a list of four general indicators that were considered representative and indicative of the selection of language resources. The indicators determine the general requirements to which the selection should be subjected. Different sets of specific criteria have been defined for each indicator. We want to emphasize again the leading role of Prof. Váradi during the discussions and in the process of the selection of the most appropriate and balanced indicators. The general indicators were, for example: a) For upgraded resources: all selected resources are state-of-the-art specimens of their type for a given language; equally valuable representatives are all included in the selection; etc.; b) For extended/linked resources: the extension of resources provides considerable value to the community, at least on regional level; the emphasis is on providing building blocks to the existing tools rather than major restructuring; etc.; c) For resources aligned across languages: no more than one tool of a certain type for each language is used; whenever applicable,

the largest set of languages is selected; etc. The general indicators were combined with the so-called Total Point Value (Maegaard, 2004) concerning the availability, quality, quantity and standards for language resource and tools under selection. Also the IPR principles and legal issues were taken into consideration, promoting the use of open data and following the Creative Commons and Open Data Commons principles.

Prof. Váradi devoted a special place in his paper to the Bulgarian contribution to CESAR project which he classified as "vital to the success of the CESAR project" (Váradi, 2014: 17). He gave the pride of place to the Bulgarian National Corpus,[3] which is "not only impressive but stands unique within the CESAR languages in terms of size and composition" (Váradi, 2014: 17). Prof. Váradi also recognised the Bulgarian WordNet; the Bulgarian Sense–annotated Corpus, Bulgarian-X Language Parallel Corpus, etc. He stressed that the Bulgarian contribution was not confined to resources unique for the Bulgarian language, but included resources made cross-lingual with the participation of all partners (Váradi, 2014: 18). It must be said that the special attention to the Bulgarian resources and tools is paid by Prof. Váradi due to the fact his paper is the key paper in a volume devoted to Bulgarian resources and tools produced within the CESAR project. Prof. Váradi has always paid equal attention to the project partners and has succeeded in an extremely delicate and appropriate way in both praising for success and in insisting on overcoming shortcomings, if any.

Prof. Váradi also described the activities of the partners in order to strengthen the position of language technologies by means of presenting the series of Language White Papers (Rehm and Uszkoreit, 2013), which describe the state-of-the-art overview of 30 European languages from the perspective of the maturity of language technologies.[4] For each of the project languages a special volume of the series was developed, where the White Papers shed light on the language technologies in Bulgarian, Croatian, Hungarian, Polish, Serbian, and Slovak in 2013, analysing criteria such as quantity, availability, quality, coverage, maturity, sustainability, and adaptability of available language resources and tools. Quantity was measured on the answers to the question: Does a tool/resource exist for the language at hand? And the more tools/resources existed, the higher the rating was. Availability was ranked by answering the question: Are tools/resources accessible, freely usable on any platform

[3] http://search.dcl.bas.bg
[4] http://www.meta-net.eu/whitepapers/overview

or only available at a high price or under very restricted conditions? Quality was assessed by answering the question: How well are the respective performance criteria of tools and quality indicators of resources met by the best available tools, applications or resources? Coverage was correlated with the answers to the questions: To what degree do the best tools meet the respective coverage criteria?; To what degree are resources representative of the target language or sublanguages? Maturity was measured with the answers to the questions: Can the tool/resource be considered mature, stable, ready for the market?; Can the best available tools/resources be used out-of-the-box or do they have to be adapted? Sustainability was related with how well the tool/resource could be maintained/integrated into current information technology systems and adaptability – with how well the best tools or resources could be adapted/extended to new tasks/domains/genres/text types/use cases, etc. and played a major role during and after the project in assessing the state of language resources for European readers. Even eight years later, the Language White Papers are important in their methodology, in the way they conduct the investigation and analyze the results, as well as in measuring the progress in creating language resources and tools for individual languages. No language was considered to have "excellent support", only English was assessed as having "good support", followed by languages such as Dutch, French, German, Italian and Spanish with "moderate support". Languages such as Bulgarian, Hungarian and Polish exhibited "fragmentary support" in 2012. As Prof. Váradi described in his paper, the Language White Papers "were used extensively by the project partners to disseminate information about META-NET and CESAR at the national level to different stakeholders, primarily through the set of CESAR roadshows, one-day high-level events each dedicated to one language" (Váradi, 2014: 19). These events represented an ideal opportunity to spread the Language White Papers as widely as possible: target users were representatives of research centres, small and large technology corporations, translation services and other users or producers of Language Technology, language communities and societies, and policy makers responsible for supporting research and innovation, economy and Information and Communication Technology. For example, the participants at the roadshow in Sofia in 2012 were over 80.

Prof. Váradi compared the available resources and tools for Bulgarian, Croatian, Hungarian, Polish, Serbian, Slovak before and after the end of the CESAR project: at the beginning of the project the initial list of tools

and resources potentially available for enhancement numbered 130 in total (Váradi, 2014: 22). According to the analysis made by Prof. Váradi, the most valuable progress was achieved in the field of multilingual and national corpora. Also, a special effort was made to render language independent resources and tools and to support cross-linguality both in the field of resources and of tools. Prof. Váradi emphasised that "progress was not only quantitative (extension, upgrade, new resources and tools), but also qualitative" (by means of intellectual property right clearance and carefully prepared and detailed resource metadata standardisation and development) (Váradi, 2014: 21). He described the number of requirements which were set up in order to meet the long-term sustainability of language resources and tools: careful selection of language resources by means of especially designed methodology; performing particular actions to ensure quality and quantity of the selected resources – upgrading, extending and linking the resources across languages; making resources visible and accessible through the META-SHARE platform and extensive metadata descriptions based on established standards.

3. The project leader

As Prof. Váradi concluded, the main role of the CESAR project was Serving Multilingual Europe. We can further generalise that the role of Prof. Váradi as project coordinator and key researcher in the successive European projects (*Multilingual Resources for CEF.AT in the legal domain* – MARCELL,[5] funded by Connecting Europe Facility / Telecommunications sector, 01.10.2018–31.03.2021 and *Curated Multilingual Language Resources for CEF AT Action* – CURLICAT[6] funded by Connecting Europe Facility / Telecommunications sector, 01.06.2020 – 31.05.2023) is invaluable. His endless and highly significant dedication to contributing to the development of powerful multilingual, cross-lingual and monolingual technologies for European languages; to facilitating the strengthening of the European language technology community, uniting industry, innovation and research; to being one of the most prominent language technology policy makers and visionaries across Europe and beyond have left an everlasting mark on the present and future of multilingual, technologically advanced Europe.

[5] https://marcell-project.eu
[6] https://curlicat.eu

References

Maegaard, B.: The NEMLAR Project on Arabic Language Resources. In: 9th EAMT Workshop, “Broadening horizons of machine translation and its applications”, 26–27 April 2004. Malta. pp. 124–128. European Association for Machine Translation (2004)

Rehm, G., Uszkoreit, H. (eds). META-NET Strategic Research Agenda for Multilingual Europe 2020. Springer, Heidelberg, New York etc. (2013)

Váradi, T.: Serving Multilingual Europe: The CESAR Project. In: Koeva. Sv. (ed.) Language Resources and Technologies for Bulgarian. pp. 9–28. “Professor Marin Drinov” Publishing House of the Bulgarian Academy of Sciences, Sofia (2014)