REVIEW

INTRODUCING THE CESAR PROJECT

Tamás Váradi*
Research Institute for Linguistics
Hungarian Academy of Sciences

This brief article is intended to introduce a recently launched EU-funded collaborative effort to enhance, extend, standardise language resources and tools from the Central and South-East European Region and make them available in an open linguistic infrastructure.

The overall aims of the project should be seen in the context of the linguistic challenges that arise from modern-day globalized digital communication. Many European languages run the risk of becoming victims of the digital age as they are underrepresented and under-resourced online. Huge regional market opportunities remain unused today because of language barriers. If we do not take action now, speaking their native language will become a social and economic disadvantage for many European citizens.

Innovative multilingual Language Technology is the ultimate intermediary that can help all European citizens to participate in an egalitarian, inclusive, and economically successful knowledge and information society. Language technology can be an enabler of instantaneous, cheap, and effortless communication and interaction across language boundaries.

From speech recognition and automatic summarisation to text mining and machine translation language technology offers ground-breaking perspectives. All these brilliant tools

* varadi@nytud.hu

and technologies are fuelled by data. The more the better and preferably integrated with linguistic knowledge in the form of annotation. This added value is what turns data into valuable resource – language resource.

The CESAR (Central and South-east europeAn Resources) project aims to contribute to a pan-European digital resource exchange facility by collecting resources and by documenting, linking and upgrading them to agreed standards and guidelines. The project will make available a comprehensive set of language resources and tools covering the Hungarian, Polish, Croatian, Serbian, Bulgarian and Slovak languages. Resources will include interoperable mono- and multilingual spoken and written databases, corpora, dictionaries and wordnets, as well as tools: tokenisers, lemmatisers, taggers, and parsers.

The CESAR project is part of a wider network of excellence called META-NET, a Network of Excellence funded by the European Union. It currently consists of 44 members, representing 31 EU countries. META-NET cooperates with a dozen other large initiatives like CLARIN, which is helping social sciences to establish the field Digital Humanities in Europe. META-NET is dedicated to fostering the technological foundations for establishing and maintaining a truly multilingual European information society that

• makes possible communication and cooperation across languages,

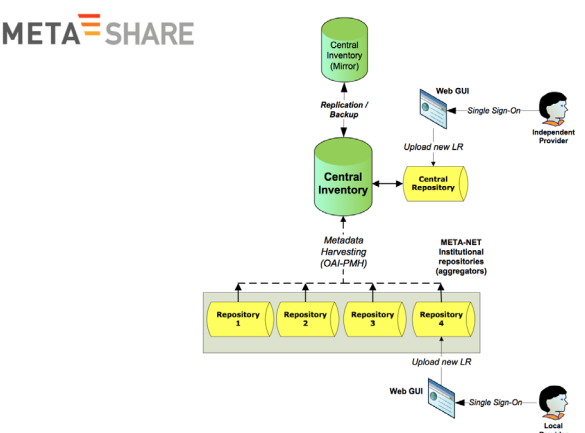
• safeguards equal access to information and knowledge for users of any language,

• offers advanced functionalities of networked information technology to all citizens at affordable costs.

The mission of META-NET is to contribute to the goal of making the multilingual European digital information space a success like written culture after Gutenberg. CESAR actively collaborates with other partner projects within META-NET, ensure consistent approaches, practices and standards aimed at ensuring a wider accessibility of and easier access and reuse of quality language resources.

The CESAR project aims to stimulate ICT-based cross-lingual communication, collaboration and participation and thereby contribute to the creation of a pan-European digital single market by stimulating ICT-based cross-lingual communication, collaboration and participation.

One of the main goals of CESAR project is to bridge the technological gap between the Central and South-east European region and the other parts of Europe by filling obvious and important gaps in language resources and tools infrastructure. ICT research has started in this region with a lag behind Western European countries. Language technology has emerged in the respective participating countries autonomously, i.e. with national support both in the academic and in the private sector. As a result, the resources developed often reflect the isolated circumstances of their creation, and still often lack standardisation. The main actors of ICT research are however now ready to reinvigorate cooperation between key technology partners in the region, and to integrate national resources on a higher level in order to make them more accessible and interoperable, making them available to the wider language technology community to ease and speed up the provision of multilingual online services. To this end, existing resources are going to be assembled and upgraded so that they comply with widely used standards or community practices.



Key resources covered by the CESAR project are going to be linked and made interoperable using the facilities of the META-SHARE repository. META-SHARE aims to build an open resource exchange infrastructure. The target user community of the resources practically embraces all stakeholders at the modern digital market: everyday end-users, professional end-users (business, administration, media, education, libraries, etc.) as well as expertise holders (researchers, industrialists, policy makers, etc). Its concern is a careful investigation of the needs of various types of users – from individual users to large multinational organisations – from the perspective of the current status as well as from the near future prospects.

The CESAR project will contribute valuable resources to META-SHARE, which will eventually be an important component of a language technology marketplace for HLT researchers and developers, language professionals (translators, interpreters, content and software localisation experts, etc.), as well as for industrial players, especially SMEs, catering for the full development cycle of HLT, from research through to innovative products and services.

One of the main goals of CESAR project is to mobilise national and regional actors, public bodies and funding agencies by organizing meetings and other focused events. The such event was the conference titled 'META-Forum 2011 – Solutions for Multilingual Europe', which was held in Budapest at the end of June. It was an official event of the Hungarian Presidency of the Council of the European Union. There were sessions about European Language Technology industries and multilingual LT in European institutions, panel discussions about multilingual information society, and invited keynote lectures by representatives from Google and IBM. More information is available at http://www.meta-net.eu/events/meta-forum-2011/index_html

CESAR project in harmony with META-NET aims to provide description materials of the national landscape of participating countries. This is the META-NET white paper series 'Languages in the European Information Society', which reports on the state of each European language with respect to Language Technology and explains the most prominent risks and chances. The series cover all official European languages. While there are numerous valuable and comprehensive scientific reviews on certain aspects of individual languages and the language technology available for them, there was as yet no generally understandable survey that summarises the main findings and challenges for each language. The META-NET white paper series was intended to fill this gap. The printed version of white paper series was distributed first time at META-Forum in Budapest.

The project started on the 1st of February and runs for two years. The resources are expected to be published in three batches. The first delivery will be made in December, 2011 and it will be followed by one in July 2012 and finally in January, 2013.

Consortium members:

– Research Institute for Linguistics, Hungarian Academy of Sciences (Hungary)
– Budapest University of Technologies (Hungary)
– University of Zagreb, Faculty of Humanities and Social Sciences (Croatia)

Institute of Computer Science, Polish Academy of Sciences (Poland)

– University of Łodz (Poland)
– Faculty of Mathematics, University of Belgrade (Serbia)
– Institute Mihajlo Pupin (Serbia)
– Institute for Bulgarian Language (Bulgaria)
– L. Štúr Institute of Linguistics (Slovakia)

Consortium coordinator:

Tamás Váradi

Research Institute for Linguistics, Hungarian Academy of Sciences

References:

Project fact-sheet
http://ec.europa.eu/information_society/apps/projects/factsheet/index.cfm?project_ref=271022

Project web-page:
http://www.meta-net.eu/projects/cesar/

META-SHARE:
http://www.meta-net.eu/meta-share