CESAR

CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES

META-NET

Project ID card

- Funded under The Information and Communication Technologies Policy Support Programme
- Call: CIP ICT-PSP-2010-4 Theme 6: Multilingual Web
- Total cost: €4.16m
- EU contribution: €2.086m
- Project reference: 271022
- Execution: from 2011-02-01 to 2013-01-31
- Project status: running
- Contract type: ICT-PSP PB Pilot Type B
- Project web page: http://www.cesar-project.net, http://www.meta-net.eu/projects/cesar

Challenge

Human language technologies crucially depend on language resources and tools that are usable, useful and available. In the last decade linguistic resources have grown rapidly for all EU languages, including lesser-resourced languages. However, even where language resources and respective tools are available they have been developed mostly in a sporadic manner, in response to specific project needs, with relatively little regard to their long-term sustainability, IPR status, interoperability, reusability in different contexts as well as to their potential deployment in multilingual applications.

Even where language resources and respective tools have been developed in sufficient quantity, they are difficult to deploy, because typically they have been idiosyncratically designed and are thus of a low interoperability level, which poses an obstacle to portability across languages, domains, and applications. It is difficult or in many cases impossible to get access to resources that are scattered around different places, are not accessible online, reside within research institutions and companies and exist as “hidden language resources”, similar to the existence of the “hidden web”.

High fragmentation and a lack of unified access to language resources are among key factors that hinder European innovation potential in language technology development and research.

Proposed solution

The CESAR project, in close harmony with META-NET and sensitive to the dynamics of community practices, intends to address this bottleneck by means of enhancing, upgrading, standardising, and cross-linking a wide variety of language resources and tools, as well as making them accessible, thereby contributing to an open linguistic infrastructure. The project will make available a comprehensive set of language resources and tools covering the Bulgarian, Croatian, Hungarian, Polish, Serbian, 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.

Project objectives

The main goals of CESAR project are:

- provide a description of the national landscape in terms of language use; language-savvy products and services, language technologies and resources; main actors (research, industry, government and society); public policies and programmes; prevailing standards and practices; current level of development, main drivers and roadblocks;
- contribute to a pan-European digital resources exchange facility by collecting resources and by documenting, linking and upgrading them to agreed standards and guidelines;
- collaborate with other partner projects, in particular concurrent ICT-PSP 6.1 pilot projects and the META-NET network of excellence – and where useful, with other relevant multi-national forums or

activities, such as FlaReNET and CLARIN – to ensure consistent approaches, practices and standards aimed at ensuring a wider accessibility of, easier access to and reuse of quality language resources and tools;

- help build and operate broad, non-commercial, community-driven, inter-connected repositories, exchanges, facilities etc. that can be used by language researchers, developers and professionals;
- mobilise national and regional stakeholders, public bodies and funding agencies by raising awareness, organizing meetings and other focussed events;
- reinvigorate cooperation between key technology partners in the region, building on previous collaboration in TELRI, MULTEXT-EAST and other projects;
- bridge the technological gap between this region and the other parts of Europe by filling obvious and important blind spots in language resources and tools infrastructure.

Target outcome and expected impact

The resources made available by the CESAR consortium are expected to be employed in complex LT applications built by initiatives of various communities in research and industry, possibly serving multiple purposes in input and intermediary modules. Since in such procedures the provided resources become further processed and structured, the extent to which they are utilized is not straightforward to estimate by figures in e.g. webservice logs, in contrast to scenarios not addressed by CESAR, such as research and education purposes where the usage of tools and datasets is measured by the number of logins and downloads.

Target Users

The target users are practically 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). Our concern is a careful investigation of the needs of various types of users – from individual users to large multinational organisations.

Cooperation

The cooperation between the partners in the consortium builds on previous joint projects such as the MULTEXT-EAST project and the TELRI initiative, as well as on their regular professional communication beyond these projects, and their collaboration, both formal and informal, on the enhancement of their resources and tools. The partners will also cooperate directly with the META-NET network, which will mainly provide the methodological, organisational, and technical foundations of a broad, distributed infrastructure.

Project partners

1. Nyelvtudomanyi Intezet, Magyar Tudomanyos Akademia (HASRIL), Budapest, Hungary
2. BUDAPESTI MUSZAKI ES GAZDASAGTUDOMANYI EGYETEM (BME), Budapest, Hungary
3. Sveučilište u Zagrebu, Filozofski fakultet – University of Zagreb, Faculty of Humanities and Social Sciences (FFZG), Zagreb, Croatia
4. Instytut Podstaw Informatyki Polskiej Akademii Nauk (IPIPAN), Warsaw, Poland
5. Uniwersytet Łódzki (ULODZ), Łódź, Poland
6. Faculty of Mathematics, University of Belgrade (UBG), Belgrade, Serbia
7. Institut Mihajlo Pupin (IPUP), Belgrade, Serbia
8. Institute for Bulgarian language Prof Lyubomir Andreychin (IBL), Sofia, Bulgaria
9. Jazykovedny ustav L'udovita Štura Slovenskej akademie vied (LSIL), Bratislava, Slovakia

Contact information

Tamás Váradi (project coordinator)
Research Institute for Linguistics
Hungarian Academy of Sciences
Benczur u 33
1068 Budapest
Hungary
Tel: +36 1 342 9372 ext. 6010
Fax: +36 1 322 9297
E-mail: varadi@nytud.hu