CESAR
CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES

PUBLISHABLE SUMMARY

Introduction

The CESAR project, in close harmony with META-NET and sensitive to the dynamics of community practices, intends to of enhance, upgrade, standardise, and cross-link a wide variety of language resources and tools, as well as making them accessible, thereby contributing to an open linguistic infrastructure. The project made available a comprehensive set of language resources and tools covering the Bulgarian, Croatian, Hungarian, Polish, Serbian, and Slovak languages. Resources include interoperable mono- and multilingual spoken and written databases, corpora, dictionaries and wordnets, as well as tools: tokenisers, lemmatisers, taggers, and parsers. During all three CESAR batches an impressive number of resources were selected, upgraded to the envisaged level and published in the META-SHARE repository, 251 all together, among them 120 corpus resources, 65 lexical conceptual resources and 66 tools and services.

	RILHAS	TMIT	FFZG	IPIPAN	ULODZ	UGB	PUPIN	IBL	LSIL	∑
Tools / Services	6	3	5	19	5	6	0	16	6	66
Corpora	19	21	12	17	11	10	0	9	21	120
Lexical/Conceptual resources	6	1	9	23	1	3	2	11	9	65
Total	**31**	**25**	**26**	**59**	**17**	**19**	**2**	**36**	**36**	**251**

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 focused 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;
- promote META-SHARE licences and maintain (for at least two years) META-SHARE nodes gathering metadata (and LRs) of the selected resources.

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.

The target users of the foreseen solution 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.

The work performed in the second project year

WP1 *(Management, M01–M24)*
All management activities (including financial, technical and risk management) started in the first year were successfully proceed in the second year.

WP2 *(Analysis and selection of language resources, M01–M18)*
Partners continued charting of the national scene of their language community landscape for further resources. This work performed covered mapping of the language service industry and language technology industry as well as local policy makers. New resources were selected for further activities. A special interest were placed to the resources and tools produced outside of the consortium. The work on Language Whitepaper were continued in the very beginning of the second year.

WP3 *(Enhancing language resources, M01–M24)*
In the second year of the project two subsequent batches of clean and reusable resources have been delivered (in July 2012 and January 2013) and made available through the open digital exchange provided by META-NET. The delivered resources have been documented according to the metadata model made available by META-SHARE. Resource descriptions have been registered at CESAR partner nodes running the most recent version of the META-SHARE application. One of the auxilliary results of the workpackage was the design, implementation and delivery of the XSLT-based environment for generation of the human-readable descriptions of resources based on the META-SHARE metadata exported to XML (using standard META-SHARE functionality).

WP4 *(Cross-national collaboration and Pilot service, M01–M24)*
The effort of this Work Package was to enhance the availability and suitability of language resources, and to provide a top-level standardized framework for their sharing. Partners took an active part in the launch of the digital resource exchange platform. The consortium and other partner projects (mainly within the META-NET consortium) cooperated between themselves (and with other EC initiatives) in works of the META-SHARE foundation. An important task of this WP was to clear the IPR and other legal issues of the chosen resources and tools, what was done in close cooperation of the other PSP projects in subsequent iteration cycles, taking into account national specialities. The other main activity within the Work Package was to take an active part in the elaboration of the metadata model, realized again in close collaboration between all concerned PSP projects and META-NET.

WP5 *(Outreach, awareness and sustainability, M01–M24)*
The main effort of this Work Package was to prepare the project sustainable beyond the end of the EU-funded phase. A special efforts were allocated to ensure the continuation and coordination of national efforts after the project's end, e.g. with promoting language research, technology, resources and applications in national circles. A massive dissemination campaign was run, enhanced mainly

through web and through the series of nationally organized high-level awareness events („road show“) that tooke place in each country of the project.

Achieved results after the second year

The main result of the project can be measured in the number of submitted deliverables and achieved milestones. The consortium made a charting of the local/national scene of language technology field, and of the local stakeholders. The consortium – with a very tight cooperation with the META-NET – prepared the English and localized versions of the Language White Paper. A careful selection and categorization of language resources and tools was made. Partners made the required upgrade and enhancement on the resources and tools of the first batch/upload. The IPR issues of the chosen resources and tools were work out. A detailed description of the metadata was prepared and uploaded to the META-SHARE server.

The dissemination efforts went according to the plans. Partners attended and organized several conferences where CESAR was introduced. Beside presentations and posters partners prepared video lectures and published papers in local and international level.

Final results

The CESAR project specifically focuses on the assembly of basic language resources for six Central and South-East European languages (Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak). The proposed set of LRs submitted in the DoW was exceed in number from 132 to 251. The CESAR project selected, upgraded and published a set of 251 language resources from the Central and East-Europe covering a wide range of corpora, lexical resources and tools. All resources were published and made accessible through the web based META-SHARE nodes. All partners set up and intended to maintain their managing nodes containing the meta-data of the selected LRs. The covered resources were are mostly promoting META-SHARE licences, which are facilitating the META identity. A special focus was placed to dissemination activities, which were aimed to the local industry and stakeholders. All partners have organized a local event for awareness raising and disseminating results of the project.

CESAR acted not only as a sovereign body, but as a part of META-NET alliance. Partners took an active part in META-NET boards and in the preparation of dissemination materials, such as the Language White Paper and the Strategic Research Agenda and the localisation of web pages and other materials.

Address of the public website

www.cesar-project.net

Address of the central CESAR META-SHARE node

http://nlp.ipipan.waw.pl/metashare/

Relevant contact details

Coordinator:

Tamás Váradi
Research Institute for Linguistics
Hungarian Academy of Sciences
Benczúr utca 33.
1068 Budapest
HUNGARY
Tel. +36 1 3214830 ext.126
Fax. +36 1 3229797
E-mail: varadi.tamas@nytud.mta.hu