CESAR

CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES

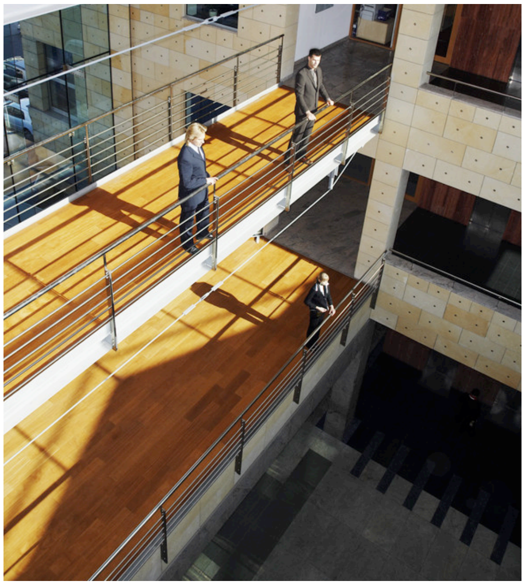
CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES

Public report

The research within the CESAR project leading to these results has received funding from the ICT Policy Support Programme (IST PSP), grant agreement no 271022.

Member of

META-NET



Many European languages run the risk of becoming victims of the digital age as they are underrepresented and under-resourced online

1 Overview

The CESAR project is an EU-funded collaborative effort to enhance, extend, standardize 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 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.

2 Aims

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 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.

- **Provide a description of the national landscape in terms of language use**
- **Mobilise national and regional actors and public bodies**
- **Reinvigorate cooperation between key technology partners in the region**
- **Bridge the technological gap between this region, by filling obvious and important gaps in language resources and tools infrastructure**
- **Contribute to a pan-European digital resource exchange facility by collecting resources**

3 CESAR and META-NET

The CESAR project is part of a wider network of excellence called META-NET, a Network of Excellence funded by the European Union.

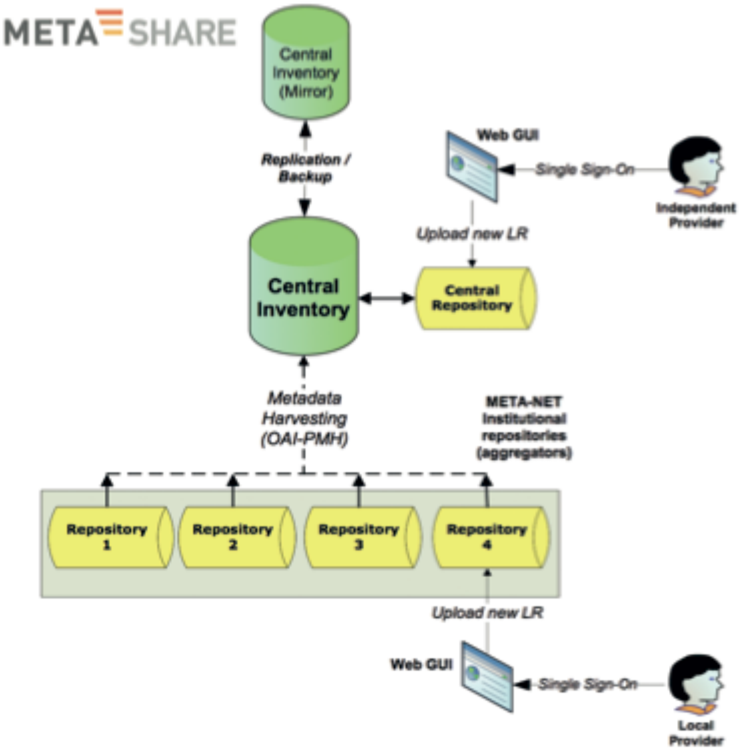
It currently consists of 54 members, representing 33 countries. META-NET cooperates with a dozen other large initiatives like CLARIN, which is helping social sciences to establish the field Digital Humanities in Europe. META-NET is dedicated to fostering the technological foundations for establishing 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 (and CESAR) 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 PSP partner projects within the META-NET framework (T4ME, METANET4U and META-NORD), ensure consistent approaches, practices and standards aimed at ensuring a wider accessibility of and easier access and reuse of quality language resources.

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 contributes valuable resources to META-SHARE, which eventually is 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.



4 Dissemination

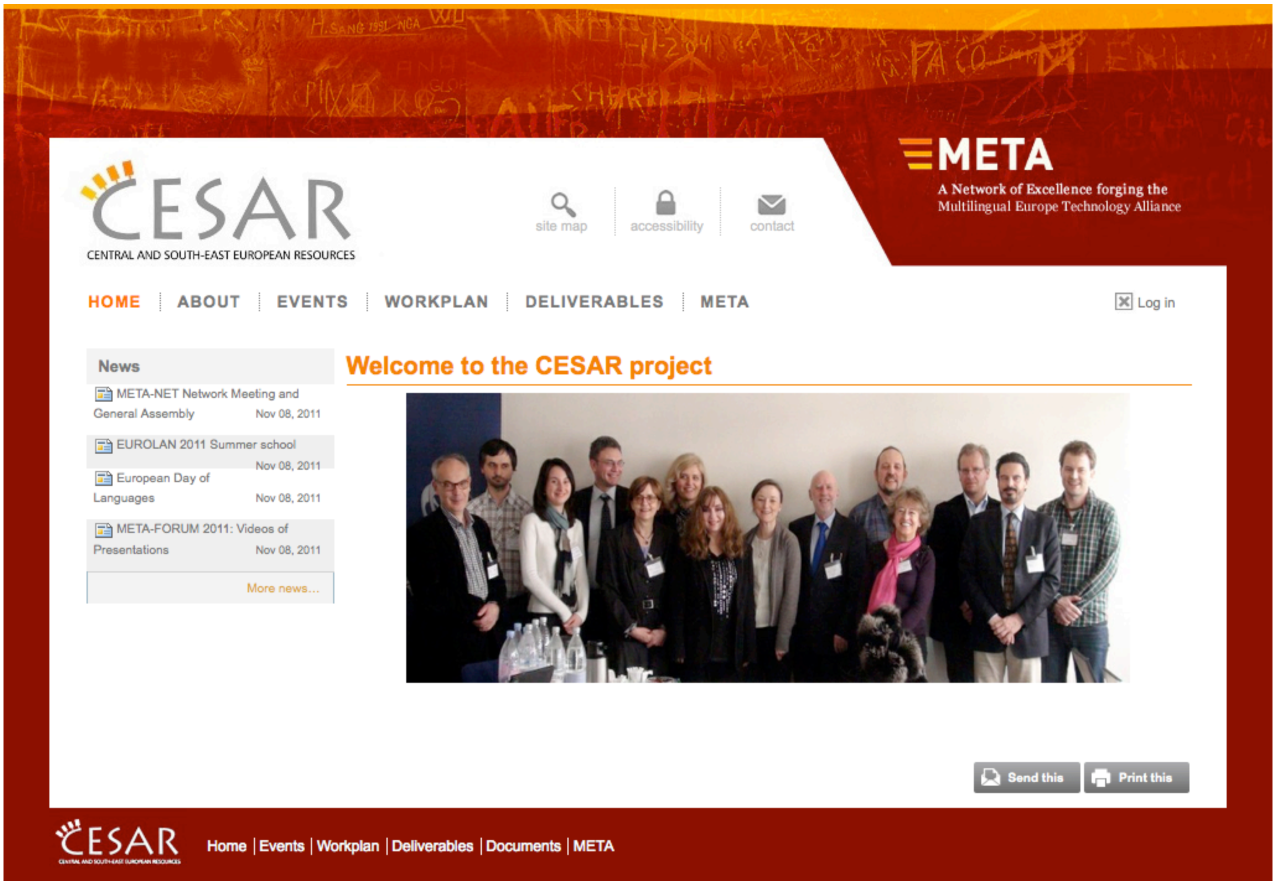
In the first months of CESAR, a communication and dissemination plan was developed in several channels on different occasions. Dissemination as well as awareness raising plays a key role within CESAR activities. Dissemination efforts are made both in local level (eg. by presentations and articles) and both in higher EU level (eg. by participating EU events and presenting its key points, by publication of CESAR activities in EU founded magazines).

CESAR activities are disseminated both by traditional and non-traditional, multi-medial channels. For dissemination purposes CESAR uses channels for visual identity (public website, paper publications and T-shirts), channels for dissemination by public appearance (participation on key conferences, participation in related projects and networks, CESAR events - like the planned road-shows, participation in professional associations).

By the channels for professionals CESAR makes its key role in LT visible by media appearance, also in printed and in electronic media. Dissemination materials (like fact-sheet, links, posters, presentations and publications) are placed in scientific and other journals, on professional portals. CESAR also has minor press releases for purpose of awareness raising, announcements and currently there is are developments on creating the Wikipedia article.

Within the modern, digital channels the CESAR web-page is playing a key role. The web page is made for non-professionals with the aim of dissemination the CESAR efforts. This effort covers articles and presentations of CESAR in an interesting form of appearance as well as video lectures taken mainly in prestigeous conferences (like the META-FORUM 2011 in Budapest).

Successful dissemination efforts were made in EU and national level by press releases such *Language Technology to tackle the Multilingual Challenge*, In: Science, Technology and Innovation. Insight into Hungarian research excellence. June 2011, p. 36–37. http://viewer.zmags.com/publication/dc678def#/dc678def/38, or *CESAR*, In: The Parliament. Politics, Policy and People, 13 June 2011., p. 71., and *CESAR, Success Stories* http://en.kpk.gov.pl/index.php?option=com_sobi2&sobi2Task=sobi2Details&catid=0&sobi2Id=218&Itemid=142&lang=en


CESAR
CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES
site map
accessibility
contact
META
A Network of Excellence forging the Multilingual Europe Technology Alliance
HOME | ABOUT | EVENTS | WORKPLAN | DELIVERABLES | META
Log in
More news...
Send this
Print this
Home | Events | Workplan | Deliverables | Documents | META

News

META-NET Network Meeting and General Assembly Nov 08, 2011

EUROLAN 2011 Summer school Nov 08, 2011

European Day of Languages Nov 08, 2011

META-FORUM 2011: Videos of Presentations Nov 08, 2011

Welcome to the CESAR project

4.1 Webpage

www.cesar-project.net

http://cesar.nytud.hu

The project webpage plays a crucial role in dissemination efforts, while it gathers all important material made for the wider publicity. The website is created to show the key points of general activity (events, news flashs, articles, video-lectures, presentations as well as project work-flow and prepared - public - deliverables) and the linkage of CESAR to the META-NET Multilingual Europe Technology Alliance.

Public web site has already been designed obeying the predefined META-NET and CESAR visual identity rules. It is technically supported and maintained by HASRIL. The domain *project-cesar.net* has been reserved and parked until the official start of the v 1.0 of the public web page. The website will also be maintained by HASRIL at least 2 years after the official end of the project.

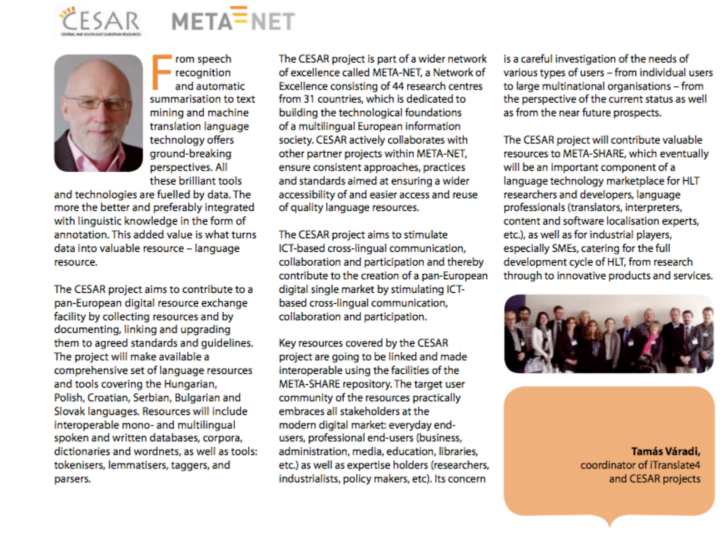
4.2 Publications

So far we have three main press releases presenting CESAR published in:

1. *Addressing the Multilingual Challenge in Digital Agenda through Language Technology*, In: The Parliament. Politics, Policy and People, 13 June 2011, p. 71. (http://cesar.nytud.hu/documents/publications/the_parliament_press_release.pdf), a magazine (see http://en.wikipedia.org/wiki/the_Parliament_Magazine) that is based upon contributions - both editorial and advertorial - from sitting members of European parliament, NGOs, pressure and interest groups on issues currently under discussion within the European institutions. Its readership and distribution covers European Parliament (all MEPs, all Secretaries-General and senior press, officers of the nine political groups, EU President and Presidency officials), Council of Ministers (senior officials in the General Secretariat of the Council), European Commission (all EU Commissioners, Chefs de Cabinet and Commission officials with responsibility for parliamentary relations), Economic and Social Committee (all members in Brussels), Committee of the Regions (Brussels secretariat), European Court of Justice (senior officials), European Investment Bank (senior officials)

2. *Language Technology to tackle the Multilingual Challenge, In: Projects Europe.* Science, Technology and Innovation. Insight into Hungarian research excellence. June 2011, p. 36–37, http://cesar.nytud.hu/documents/publications/projects_magazine.pdf,
also
http://viewer.zmags.com/publication/dc678def#/dc678def/38, a leading website and magazine (http://www.projects.eu.com/) dedicated to research and development, science, technology and innovation across Europe.

3. CESAR Success Story. In: Portal National Contact Point in Poland (http://en.kpk.gov.pl/index.php?option=com_sobi2&sobi2Task=sobi2Details&catid=0&sobi2Id=218&Itemid=142&lang=en).



CESAR META-NET

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

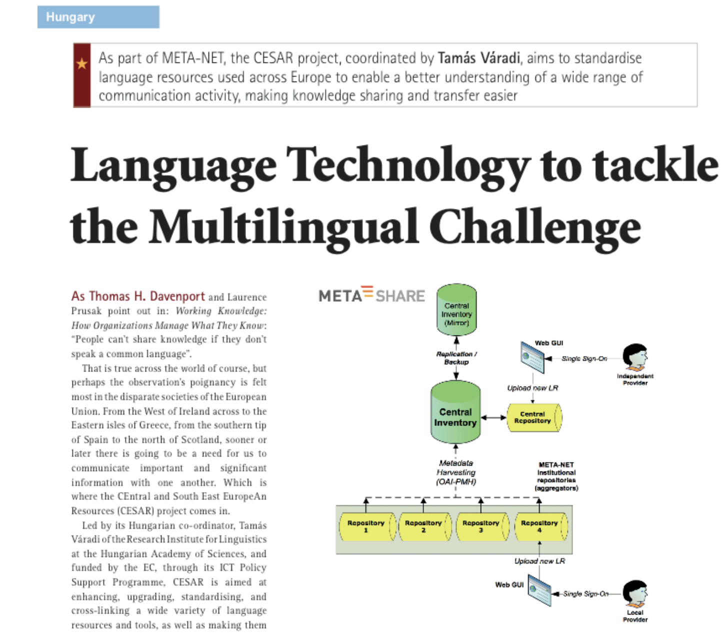
The CESAR project is part of a wider network of excellence called META-NET, a Network of Excellence consisting of 44 research centres from 31 countries, which is dedicated to building the technological foundations of a multilingual European information society. CESAR actively collaborates with other 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.

Key resources covered by the CESAR project are going to be linked and made interoperable using the facilities of the META-SHARE repository. 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 eventually will 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.

Tamás Váradi,
coordinator of iTranslate4
and CESAR projects



Hungary

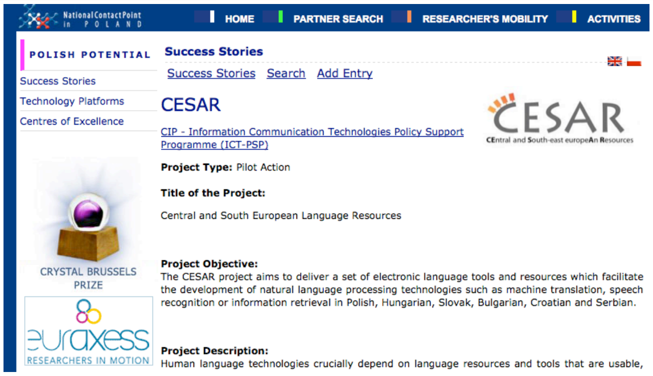
As part of META-NET, the CESAR project, coordinated by **Tamás Váradi**, aims to standardise language resources used across Europe to enable a better understanding of a wide range of communication activity, making knowledge sharing and transfer easier

Language Technology to tackle the Multilingual Challenge

As Thomas H. Davenport and Laurence Prusak point out in: *Working Knowledge: How Organizations Manage What They Know*: "People can't share knowledge if they don't speak a common language".

That is true across the world of course, but perhaps the observation's poignancy is felt most in the disparate societies of the European Union. From the West of Ireland across to the Eastern isles of Greece, from the southern tip of Spain to the north of Scotland, sooner or later there is going to be a need for us to communicate important and significant information with one another. Which is where the CEntral and South East EuropeAn Resources (CESAR) project comes in.

Led by its Hungarian co-ordinator, Tamás Váradi of the Research Institute for Linguistics at the Hungarian Academy of Sciences, and funded by the EC, through its ICT Policy Support Programme, CESAR is aimed at enhancing, upgrading, standardising, and cross-linking a wide variety of language resources and tools, as well as making them



Success Stories

Success Stories Search Add Entry

CESAR

CIP - Information Communication Technologies Policy Support Programme (ICT-PSP)

CESAR CEntral and South-east europeAn Resources

Project Type: Pilot Action

Title of the Project:

Central and South European Language Resources

Project Objective:
The CESAR project aims to deliver a set of electronic language tools and resources which facilitate the development of natural language processing technologies such as machine translation, speech recognition or information retrieval in Polish, Hungarian, Slovak, Bulgarian, Croatian and Serbian.

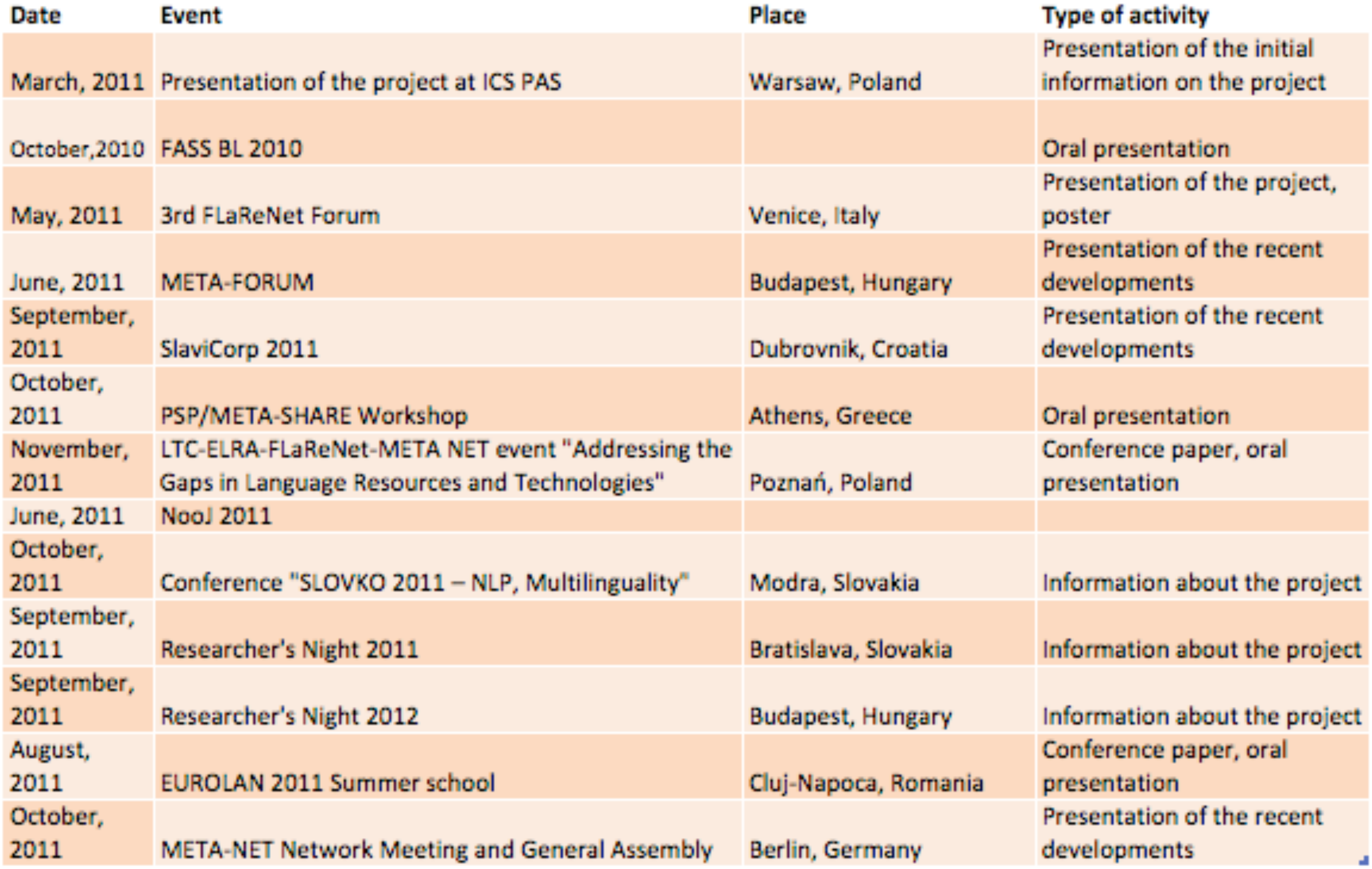
Project Description:
Human language technologies crucially depend on language resources and tools that are usable,

4.3 Events

Events – internal meetings, small workshops and also large, public conferences – play an important role in CESAR. In the reporting period representatives of CESAR participated in multiple events (within the META-NET alliance). In addition, CESAR organised or co-organised several public and also internal events.

The smaller events of CESAR activity are internal conferences, presentations on smaller conferences and workshops on national level (eg. European Day of Languages, EUROLAN 2011 Summer School).

The activity of CESAR is presented in several key LT conferences an workshops with purpose of dissemination our main efforts and aims. The schedule of main events of the first period of the project are listed in the next table.



Date	Event	Place	Type of activity
March, 2011	Presentation of the project at ICS PAS	Warsaw, Poland	Presentation of the initial information on the project
October,2010	FASS BL 2010		Oral presentation
May, 2011	3rd FLaReNet Forum	Venice, Italy	Presentation of the project, poster
June, 2011	META-FORUM	Budapest, Hungary	Presentation of the recent developments
September, 2011	SlaviCorp 2011	Dubrovnik, Croatia	Presentation of the recent developments
October, 2011	PSP/META-SHARE Workshop	Athens, Greece	Oral presentation
November, 2011	LTC-ELRA-FLaReNet-META NET event "Addressing the Gaps in Language Resources and Technologies"	Poznań, Poland	Conference paper, oral presentation
June, 2011	NooJ 2011		
October, 2011	Conference "SLOVKO 2011 – NLP, Multilinguality"	Modra, Slovakia	Information about the project
September, 2011	Researcher's Night 2011	Bratislava, Slovakia	Information about the project
September, 2011	Researcher's Night 2012	Budapest, Hungary	Information about the project
August, 2011	EUROLAN 2011 Summer school	Cluj-Napoca, Romania	Conference paper, oral presentation
October, 2011	META-NET Network Meeting and General Assembly	Berlin, Germany	Presentation of the recent developments

5 Results

In the first period of the project (until the publication of the first batch of selected resources) the following results had been achieved.

For the further dissemination activity efforts CESAR partners have charted of the national scene of their language community landscape. There are several channels of aiming the CESAR objectives, therefore partners made mapping of several types of potencional target groups. The charting of the national scenes had influence on language service industry and language technology industry, as well as on the local policy makers in their country.

In important work was the identification and detailed description of national language resources which are already developed or currently under development. For further and for more detailed description of resources an on-line questionnaire was elaborated, made specially for catalogization of the identified language resources and tools. Together with the data of the on-line questionnaire a report on available and potentially available resources was prepared.

An important role of national charting was the preparation of the Language Whitepapers which ran in cooperation with other META-NET members. The Whitepaper series based on the exhausting effort of partners in involving and describing the national language technology scene and describing the chosen resources and tools. The participation in preparing the Language Whitepapers resulted Languages Whitepapers (description of the language situation) for six countries (Bulgaria, Croatia, Hungary, Poland, Serbia, Slovakia) both in English and national languages.

Other basic activity of the CESAR project is the continuous enhancement of the chosen language resources. It covers an upgrade of metadata related materials on resources and tools with tight cooperation with META-SHARE. With the help of META-SHARE there were created several license templates and defined basic principles according to the partners have selected their resources into the META-SHARE. There was created a metadata description model architecture and several description schemes for the chosen resources. Partners carefully categorized their resources for further operation (conversion/upgrade/enhancement activities according to the implementation plan) and have prepared a selection of language resources.

Open source format, cross linking and aligning is one of the basic achievements of the CESAR project. Due to this reason there have been launched an upgrading the open source language tool NooJ (mainly initial activities on translating NooJ to open source platform).

6 Consortium members

Participant no.	Participant organization name	Participant short name	Country
1 (CO)	NyelvtudományiIntézet, Magyar Tudományos Akadémia	HASRIL	Hungary
2	Budapesti Műszaki és Gazdaságtudományi Egyetem	BME-TMIT	Hungary
3	Sveuciliste u Zagrebu, Filozofski Fakultet – University of Zagredb, Fakculty of Humanities and Social Sciences	FFZG	Croatia
4	Instytut Podstaw Informatyki Polskej Akademii Nauk	IPIPAN	Poland
5	Uniwersytet Lodzki	Ulodz	Poland
6	Fakulty of Mathematics, University of Belgrade	UBG	Serbia
7	Institut Mihajlo Pupin	IPUP	Serbia
8	The Institute for Bulgarian Language Prof. Lyubomir Andreychin	IBL	Bulgaria
9	Jazykovedny Ústav Ludovíta Stúra Slovenskej Akadémie Vied	LSIL	Slovakia

Contacts:

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