PROJECT FINAL REPORT

CESAR

CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES

Grant Agreement number:	**271022**
Project acronym:	**CESAR**
Project title:	**Central and South-East European Resources**
Funding Scheme:	**CIP ICT PSP**
Period covered:	**01/02/2011 – 31/01/2013**

Name of the scientific representative of the project's co-ordinator[1],

Title and Organisation:	**Research Institute for Linguistics, Hungarian Academy of Sciences**
Tel:	**+36 1 3214830 ext. 126**
Fax:	**+36 1 3229797**
E-mail:	**varadi.tamas@nytud.mta.hu**
Project website address:	**www.cesar-project.net**

[1] Usually the contact person of the coordinator as specified in Art. 8.1. of the Grant Agreement.

Executive summary

The CESAR project, in close harmony with META-NET and sensitive to the dynamics of community practices, intends to enhance, upgrade, standardise, and cross-link a wide variety of language resources and tools, as well as making them accessible through an open linguistic infrastructure. The project makes available a comprehensive set of language resources and tools covering the Bulgarian, Croatian, Hungarian, Polish, Serbian, and Slovak languages. Resources include mono- and multilingual spoken and written databases, corpora, dictionaries and wordnets, as well as tools: tokenisers, lemmatisers, taggers, and parsers. During the lifetime of the project an impressive number of resources , 251 in total, were selected, upgraded to the envisaged level and published in the META-SHARE repository including 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**

Table 1. Language resources of CESAR

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 directly and in 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 of 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 was a careful investigation of the needs of various types of users – from individual users to large multinational organisations.

The project specifically focuses on the assembly of basic language resources for six Central and South-East European languages. The proposed set of LRs specified in the DoW numbered 132. 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 are mostly promoting META-SHARE licences, which are facilitating the META identity. Special attention was placed on dissemination activities, which were aimed at the local stakeholders.

CESAR META-SHARE nodes are established for long-term maintenance of the selected resources. The CESAR-partners are committed to hosting and making available the selected language resources and operating the repository of LRs for at least 24 months after the termination of the project. Within this activity all partners are committed to giving user-support, software-based and/or human services for the resources.

Summary description of project context and objectives

The main goals of the CESAR project are summarized as following:

- to provide a description of the national (resp. language community) 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; and synthesize this in a simple, clear, standardized format;
- 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;
- to collaborate with other partner projects, in particular other CIP 6.1 pilot projects within 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 and easier access and reuse of quality language resources;
- to 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;
- to mobilise national and regional actors, public bodies and funding agencies by raising awareness, organizing meetings and other focused events;
- to reinvigorate cooperation between key technology partners in the region, building on previous collaboration in TELRI, MULTEXT-EAST and other projects;
- to bridge the technological gap between this region and the other parts of Europe by filling obvious and important gaps in language resources and tools infrastructure.

The main goal of the CESAR pilot project is to stimulate ICT-based cross-lingual communication, collaboration and participation and thereby contribute to the creation of a pan-European digital single market..

The main actors of ICT research are 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 assembled and upgraded so that they comply with widely used standards or community practices. (Re)usability and interoperability is further increased by focusing on resources that can be used in multilingual or cross-lingual applications, linking chosen resources across two or more languages, bringing multilingual resources into focus, and using monolingual resources as initial building blocks in multilingual or cross-lingual applications. Key resources are linked and made interoperable using the facets in the META-SHARE repository.

The target user community of the resources practically embraces all stakeholders of 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). The concern of the project was a careful investigation of the needs of various types of users – from individual users to large multinational organizations – from the perspective of the current status as well as from the near future prospects.

The accepted philosophy is to turn from the language-related community towards the digital market as a whole, mediating between the two, providing highly reliable market evaluation by means of designing different groups of end-users, evaluating usage, impact and potential adoption of LRTs, and providing multidisciplinary and cross-lingual observations.

LRT accessible in this way ensure access to recently updated versions, mechanisms for authorisation and authentification, providing control ranging from free to individually restricted access.

Maturity of the technical solution

The starting point for our work is that the fundamentals for achieving the objectives set by CESAR have already been laid down by the participating countries individually. This however also means that existing resources had emerged in the respective language communities autonomously before the CESAR project came into being, whereas the current state-of-the-art already would require and allow for improving over such baseline situation. In each country in the past, various research communities in academia and industry started compiling either general-purpose corpora, or, at the other end of the spectrum, narrowly focused resources, such as training corpora for specific purposes (e.g. speech synthesis).

Orthogonal to this, as a general property, available tools and data tend to feature analogous characteristics across languages, since they came into being based on the actual state-of-the-art research foci in computational linguistics. For example, typically there exist for each language a literary corpus, a finances/news domain corpus, a web corpus, collections based on Wikipedia, compilations of biomedical texts and of emotion research, and so on. Before the CESAR project there exeisted no complex large scale applications either at the national levels, or across borders or language communities, but instead many of the resources were made ready for straightforward integration.

The CESAR project brought LRT of respective languages to the same desired technological level (accessibility/availability using web-services where appropriate, interoperability based on common standards and common metadata formats) and made them accessable through META-SHARE not just for research but also business-oriented purposes. Publicly funded research communities as well as business communities (or IT SMEs) that develop applications for the market are potential users of the CESAR project outcomes. The project partners not just predicted the directions and pace of LRT development in their respective countries, but they also shape future directions and applications with their own activity. In this respect CESAR do not just follow the predictions but is taking part in paving the way for these languages to the European single digital market.

Target outcome and expected impact

The CESAR project specifically focuses on the assembly of basic language resources for six Central and South-East European languages, all of them considered, by any comparison, less-resourced: four of them (Hungarian, Polish, Bulgarian, Slovak) being official languages of recently joined member states, while two languages (Croatian and Serbian) represent languages of states scheduled to join the EU in the near future. The coverage of these languages brings about an added benefit of the project, anticipating and meeting foreseeable requirements with respect to resources from these languages. Building on a wide range of already existing resources and previous national or international activities, the project created, populated and operated a comprehensive language-resource platform enabling and supporting large-scale multi- and cross-lingual products and services. In extensive cooperation with other similar EU initiatives, resources in the CESAR languages were upgraded and updated to widely acknowledged standards, thus ensuring interoperability and creating the ground for widespread and efficient accessibility and the potential to modularize them in language technology pipelines.

The resources made available by the CESAR consortium are expected to be employed in complex LT applications built by joint initiatives of various communities in research and industry, possibly serving multiple purposes in input and intermediary modules. Since in such procedures the provided resources might 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 popularity of tools and datasets is possible to measure by the number of logins and downloads.

All consortium partners in the CESAR project are leading national centres with expertise in resource development and in cooperation at national as well as international level. The partners' ability to mobilize their national language communities has been demonstrated in several joint efforts in creating national linguistic infrastructures. In many of the partners countries there was already an established framework of support from public authorities and agencies. The CESAR project intended to maintain strong ties with partner projects with respect to harmonising strategies and synchronising activities. In addition to regular contacts, this commitment was reinforced by the invitation of META-

NET to representatives of these projects to attend the respective policy making boards. The involvement of the CESAR and partner projects in the common policy making process serves not only the purpose of ensuring that the project output will be maximally compliant with the technical requirement of the META-NET but will also create new synergies among the partner projects themselves – not only in the lifetime of the projetc(s), but perspectively for a longer time. This aim is now underlined with the commitment of partners to maintain META-SHARE nodes for at least two years after the project end.

Long term viability

The resources integrated within the CESAR project constitute important building blocks of language processing research and technology systems; for their long term sustainability they must receive long-term support after the termination of the project. Regarded as crucial for both the language (technology) community and the future participants of a single digital market, the resulting standardised tools and resources are made available either via servers of the original creators and handed over to META-SHARE nodes set up in each country covered by the project. It is envisaged that the resources will also be made available through other collaborative platforms such as those developed in the FLaReNET and CLARIN projects, or dissemination channels of related user communities such as digital humanities and digital libraries.

The major aspects of sustainability of the project are briefly listed below:

- Maintenance: The constant maintenance of resources delivered by the project is the minimum requirement for their preservation. The internal maintenance (e.g. provided by the consortium) of project outcome should be carried out independently of its external applications.
- Development: Services of the CESAR platform are constantly improved to support user requirements.
- Documentation: Based on resources documentation, guidelines are produced to provide first level support to repository users. Apart from descriptions of tools and resources, their scope, as well as input and output methods, representative usage examples should be described.
- User support: The guarantee of users' satisfaction to get satisfactory support is an important measure to address in long-term maintenance. The level of provided support will be carefully planned to maintain realistic response time and avoid unnecessary overhead.
- Standardization and interoperability: Standardization and interoperability issues are the core of the project and they form a prerequisite for the project success, to be detailed below.

The above-mentioned aspects have been discussed by the consortium members in order to indicate clear-cut criteria for measuring project sustainability (during and after the project), driven by the following considerations:

During the CESAR project maintenance of resources was assured, representing one of the project’s core activities. Resources that have been developed so far are basic building blocks of language technology for each consortium partner in its respective country and language. These resources and tools have been built primarily out of national and partly out of EU funding. Their upgrade, cross-linking and bringing to widely accepted standards was carried out in CESAR. Such large-scale pre-competitive resources have been only marginally funded by industrial partners.

The long term viability of the chosen resources is stressed by the following points:

- all chosen LRs are presenting a high value in the respective language community (either are valuable for their high level of professionalism or are representing a valuable resource made available in a range of linked languages);
- all chosen resources were subject to a well organized and prepared upgrading process which clearly shows the quality and usability both from industrial and R&D side.

In order to facilitate log term viability of the solution laid down in the project and the outcome itself (the 251 pieces of LRs) participants set up the following META-SHARE nodes, which will be maintained by the hosting institutes for at least two years after the project end:

- RIL MTA node: http://metashare.nytud.hu/,
- BME-TMIT node: http://metashare.tmit.bme.hu/,
- FFZG node: http://meta-share.ffzg.hr/,
- IPIPAN node: http://nlp.ipipan.waw.pl/metashare/,

- ULodz node:http://metashare.ia.uni.lodz.pl/,
- IBL node: http://metashare.ibl.bas.bg/,
- UBG node: http://meta-share.matf.bg.ac.rs,
- LSIL node: https://metashare.korpus.sk/

META-SHARE nodes

For batches 1 and 2, Partners agreed to set up one official META-SHARE node in Warsaw, Poland, hosted and maintained by IPIPAN and mirrored at ULodz. At the end of the project META-SHARE nodes were organized into a hierarchical structure: managing nodes are synchronized, and provide all META-SHARE metadata and resources, whilst network nodes are not synchronized, but automatically harvested by a managing node.

As the META-SHARE server software runs in full functionality (including synchronization), CESAR Partners decided to promote the original Warsaw node to become a managing node (CESAR managing node), and set up a network node at each partner's premises to provide metadata for harvesting by the CESAR managing node, which shares metadata with other META-SHARE managing nodes.

CESAR META-SHARE nodes are committed to long-term maintenance of the selected resources. CESAR-partners committed to host and make available the selected language resources and host the repository of LRs for at least 24 months after the termination of the project. Within this activity all partners agreed to give user-support, software-based and/or human services and start and continuing in the META-SHARE software development team (requiring Python and Django skills).

A description of the main S&T results/foregrounds

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

The participants of the project are key players of ICT research in their respective country and EU wide dimensions. Their effort in this project is 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 maintained either by the partners of the consortium or by third parties were assembled and upgraded so that they comply with widely used standards or community practices. (Re)usability and interoperability were in the target of the focus with the aim to use resources in multilingual or crosslingual applications, linking these resources across two or more languages, bringing multilingual resources into focus, and using monolingual resources as initial building blocks for furthet processing – mainly as parts of various NLP applications. Key resources were linked and made interoperable using the facets in the META-SHARE repository.

One of the promoted philosophies of the project is to turn from the language-related community towards the digital market as a whole, mediating between the two, providing highly reliable market evaluation by means of designing different groups of end-users, evaluating usage, impact and potential adoption of LRTs, and providing multidisciplinary and cross-lingual observations. This aim however had to be achieved by ful professionalism of the partners, which resulted in a high number of LRs made available through the META-SHARE repository.

The tight and fruitful cooperation between the partners in the consortium was built 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.

CESAR is an active part of META-NET alliance, promoting its main ideas and acting as member of several META-NET boards.

The primary outcomes of the project are listed as follows:

- compilation of the Language White Paper
- construction of a contact database for all participating countries involving the key players from business, development, politicians and media
- an enhanced method and set of criteria for selection of resources
- set of LRs provided and uploaded in three batches (with a detailed description of the work done on resources)
- cleared and agreed IPR descriptions and licences
- sustainability plan for the CESAR resources
- a detailed plan for awareness, mobilisation and dissemination actions
- sustainability plan for the outcomes of the project
- European-wide dissemination actions
- set up and maintenance of 6 META-SHARE nodes
-

Language White Paper

The 31 volumes of the META-NET Language White Paper series “Languages in the European Information Society” reported on the Language Technology (LT) development for 30 European languages. The Language White Paper documents sketch the main findings and challenges facing the LT also for Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak, while reviewing the most urgent risks and chances in the field. The volumes focus on the readiness of core technologies for the respective languages.

The first chapter *A Risk for Our Languages and a Challenge for Language Technology* informs about some of the risks for the European languages after the digital revolution that has changed communication and society. It also lays down a number of challenges facing the implementation of the LT tools and resources, such as language borders, multilingualism, and the slow pace of technological progress. It further explains language technology as a key enabling technology that helps people collaborate, conduct business, share knowledge and participate in society across different languages. The analysis reviews the LT application fields, such as automatic translation, content production, information processing and knowledge management. LT is a tremendous opportunity for the EU citizens to communicate across the language borders on the European common market and global market. The section further explains how computers handle language, taking a brief look at the way humans acquire languages, and then considering how machine translation systems work.

The second chapter *Bulgarian/Croatian/Hungarian/Polish/Serbian/Slovak in the European Information Society* presents in brief a couple of general facts about the respective language, including features of orthography, morphology (rich inflectional and derivational systems; aspectual verb pairs), syntax (pro-drop, relatively free word order) that may impede the development of the LT tools and resources. Further, it outlines the recent changes in Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak such as internationalization (mostly involving English loan words). The documents discuss the status of Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak as official languages in Bulgaria, Croatia, Hungary, Poland, Serbia and Slovakia and administrative languages in the EU, as well as their inclusion in the respective educational systems and curricula (at elementary, secondary and university level). A further view extends to the Internet resources in Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak (news portals, Wikipedia, web search and online translation services).

The third chapter *Language Technology Support for Bulgarian/Croatian/Hungarian/ Polish/Serbian/Slovak* addresses the state-of-art of the LT support for Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak. First, it gives a brief account on the LT application architectures that consist of several components to mirror different aspects of language. Second, it covers the situation in the LT research and education. Third, it concludes with an overview of the past and ongoing research programs in universities and institutions. The section ends with an expert estimation of the situation with core LT tools and resources. Spelling checkers and grammar checkers are available for some of the languages, although their level of precision is still unsatisfactory. A number of applications for web search are also presented reviewing their principles, merits and shortcomings. However, for a more sophisticated request for information, integrating deeper linguistic knowledge is essential. Consequently, the next generation of search engines will involve much more sophisticated language technology. It may involve subfields such as machine-reading thesauri and ontological language resources, information retrieval, named-entity recognition, among others. Retrieving relevant answers requires an analysis of the search expresson at the syntactic and semantic level as well as the availability of an index that allows for the fast retrieval of relevant documents. The processed query needs to be matched against a huge amount of unstructured data. Matching a query to documents written in a different language (the so-called cross-lingual information retrieval), is even more demanding.

The analysis proceeds to speech interaction technologies, such as voice user interfaces and text-to-speech applications, for the Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak. In this domain, a genuine market for the linguistic core technologies for syntactic and semantic analysis does in general not exist yet. Another LT application field is machine translation that is particularly challenging for all languages. For Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak question answering, information extraction, and summarization (text generation or summary generation) have not been the center of numerous initiatives.

The section also addresses the LT hidden principles. Building language technology applications involves a range of subtasks that do not always surface at the level of interaction with the user but provide significant hidden functionalities.

The White Papers shed light on the LT in Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak education including subjects and curricula, mostly at university level. Further, the White Papers mention in brief the various programs and initiatives that fund the development of the LT tools and resources for languages in question.

The documents show the results of a survey of the state-of-art of LT tools and resources for Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak in comparison with other languages covered by the META-NET initiative analysing criteria such as quantity, availability, quality, coverage, maturity, sustainability, and adaptability. The results are included into tables of tools and resources for respective languages.

The fourth chapter *About META-NET* contains an overview of the META-NET tasks, initiatives, and member organisations.

Contact database of the stakeholders

The successful awareness raising and dissemination campaign can be made only if they reach the right persons. Since one of the main pillars are several dissemination activities, partners prepared a carefully selected list of target audience, built of national and international bodies and relevant stakeholders. The database was built through the lifetime of the project and was used for several actions and was submitted as deliverable at M4 and M18.

The aim of the contact database was to present the relevant contact information for three main groups of stakeholders: Natural Language Processing research and business groups; Natural language Processing profiting groups; and public authorities. The Contact database covers individuals (experts), institutions (research, national-funding agencies, government) and companies (producers and important users) dealing with Language Resources and Technologies. Since the partners are from Bulgaria, Croatia, Hungary, Poland, Serbia and Slovakia, the contact database presents information for these countries and respective languages: Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak.

The Contact database consists of contacts information (name, affiliation and email) on the national (and international) level of representatives of the following stakeholder types: research, education, government, industry (vendors, service providers and users), associations, integrators, and others (funding agencies, media, non-governmental organisations, etc.). There is also information for the main area of the expertise, namely: Natural Language Processing (with specifying sub-domains such as Machine Translation, Speech Processing, Information Technologies), speech, translation, localisation, interpretations and others. The research communities in the field of the Natural Language Processing from the respective countries are represented by means of the most prominent and leading centres and scholars in the field working in different academic institutions, universities and private bodies. The Language service industry and Language technology industry (local translation, localization and interpretation) are represented by means of the leading companies in the respective countries and abroad. Finally, the Policy makers: politicians, administration, media, funding agencies, all affecting the language-related community and digital market, are surveyed and indicated.

To collect this information, previous surveys were taken into account, as well as the information from the internet and personal contacts were exploited. As a result a comprehensive list of contacts was collected.

Since the CESAR aims at encompassing a large variety of language resources, language processing and annotation tools and technologies, our target users in this respect are developers and researchers in both industry and academia. This includes private and public institutions, companies and individuals involved in HLT research and development: industrial organizations and SMEs, academic institutions, research organizations, universities, individual researchers and students, national governments, EU institutions, and private investors. We have targeted all these users throughout a variety of means – e-mails and personal letters, publications in different media, and a Road show of European language technology, aiming at presenting state-of-the-art, directions and visions of development of language resources and tools for the common scientific and commercial market.

Each of the events made within the Road shows (held in different countries) intended to promote knowledge about language technology and its potential, together with the possible challenges implicit in its development. Each time particular attention was drawn to language technology for the language of the host country by presenting new advances in the field and their application in administration and business as well as involvement of open source and research community in the process of development of language resources. This was achieved by gathering experts who tried to answer important questions on the future of language and language processing in a globalized digital information society.

At M18 newly collected information were introduced, on whether the specified contact has been familiarized with the Cesar project and META-NET, whether he/she has expressed interest in Cesar and META-NET, and whether one or more personal meetings with Cesar partners have been achieved.

The Contact database comprising the representatives from the different stakeholders groups is the result of our extensive efforts towards the promotion of our work within the CESAR project to all relevant groups within the community.

The Contact database may be updated and enlarged after the end of the project.

TOTAL	D2.2a (month 4th)	D2.2b (month 18th)
Contacts collected	655	870
Unique organisations identified	340	394
Distribution by area of expertise:		
NLP	151	171
Speech	12	15
Translation	18	18
Localisation	20	22
Interpretation	66	66
Other (incl. IT and ICT specialists)	388	544
Distribution by sector:		
Industry vendors	61	69
Industry language service providers	20	32
Industry corporate users	43	62
Education	176	241
Research	92	136
Government	116	154
Association	24	30
Integrator	53	73
Other	67	67
Contacts introduced to Cesar and Meta-Net	xx	600
Contacts who have expressed interest with Cesar and Meta-Net	xx	399
Contacts who have had personal meetings with Cesar partners	xx	349

Table 2. Summary of the contact database made within the CESAR project

The general statistics shows that during the reported period there were collected 870 contacts from 394 different organisations. 600 contacts were introduced to CESAR and META-Net and 399 of them have expressed interest with CESAR and META-NET.

Report on methodology and criteria for the selection of resources

The Report on methodology and criteria for the selection of resources reported on the results from the accomplishment of the Task 2.3 Selection of resources of further interest. The report describes the methodology and criteria that were used for a precise selection of resources and tools.

Methodology and criteria that allow partners to assess the quality and importance of language resources and tools were established in the starting phase of the project, thus enabling the CESAR project partners to set appropriate priorities. The aim of the report was to ensure a balanced coverage of resources and tools for different end-users and tasks, groups of products and services.

On the basis of the agreed methodology and criteria, the consortium selected the best possible mix of resources that will further raise the interest of different groups of end-users. The outcome of this task allowed CESAR partners to provide an analysis of the current situation and made suggestions in case of lack of essential resources for covered languages to determine further efforts of the community.

The report introduced the adopted methodology and criteria for language resource selection. Criteria of quality assessment were proposed with particular attention to basic developments. The approach for language resources and tools selection was based on a list of indicators, where each language resource is specified according to different groups of criteria. The goal was to ensure as accurate measurement as possible for different quality and quantity parameters.

The adopted methodology and criteria were not equally applied for each individual language, as the situation of LRs were not equal in covered languages. The criteria and methodology set up for the selection of LRs were made with aim to facilitate all covered languages with common LRs in order to reach a high level of quality and attain cross-linking and reusability (in case of tools) between them.

The report gave at time of publishing an overview of the existing language resources for the every language and made it possible to identify the gaps in the provision of the language resources and tools components. For each language, an analysis of the set of selected resources and tools was provided, together with an outline of the gaps. The analysis lead to the conclusion as to what kind of resources had to be chosen at the greatest interest in the rounds of selection.

The methodology and criteria adopted for resource selection

The first step was to develop a methodology by which the identified language resources might be evaluated. A query was distributed among the partners to solicit suggestions on how to approach the evaluation procedure. It was confirmed that no single current methodology can be accepted as a standard. Instead, the consortium developed a list of four general indicators that were considered representative and indicative for 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. The indicators are as follows:

General evaluation of resources

In this indicator, the process of enhancing the resources and tools was carried out in three flows: resource upgrade, extension, and cross-lingual alignment. Among these indicators, further classification was made with respect to the following criteria, which were taken into consideration in the process of evaluation as yes-no questions (in the text below the priorities of the selection are highlighted with bold):

For upgraded resources:

- All selected resources are state-of-the-art representatives of their type for a given language.
- Equally valuable representatives are all included in the selection.
- Current status of resources have superior quality at least on regional level without the need of excessive further development.
- Licensing issues allow free processing and access to resources and resource-related materials or the consortium succeeds in reaching an agreement with respective copyright holders.

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.
- Additional resources are integrated with the existing ones only if they significantly improve the quality of resulting resources.
- If more than one representative of a certain resource type for a language has been selected, they are very likely to be interlinked to benefit from strong sides of both solutions.
- If less-developed, but still very popular resources can benefit from the enhancement due to their well-developed equivalent, their enhancement is also considered.
- Experience of other consortium members/other consortia is extensively used in the process of extension of national resources to provide strong foundation for cross-lingual coverage.
- Tools that are language-neutral or cross-lingual, are preferred.

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.
- Language Processing Tools in NooJ.

- Language-independence is targeted to a great extent.
- The quality of a result is of immense concern.

The soundness of specification couldn’t be judged without knowing the broader context of usage, adequacy of a certain language resource. To estimate the quality, quantity and importance, every case was thoroughly examined, taking into account regional determinants, popularity of the format outside its home institution. This indicator required a complex assessment of language resources in the context of the whole set of the established criteria. The partners not only appraised whether the selected resources have fulfilled the established criteria but also provided concrete examples and detailed explanations based on thorough analysis.

Total Point Value

Following the approach of the EU NEMLAR (Network for Euro-Mediterranean Language Resources) project (concerning a BLARK for Arabic), the notions of availability, quality, quantity and standards were further specified and taken into account in the process of language resource selection. A technique, supplementing the NEMLAR approach, while defining exact measures for quality and quantity aspects and incorporating the standardisation into the quality section, was developed at the first period of the project. The evaluation process consisted of the following steps: specification of the point value (PV) of every measure for each resource; aggregation of the points into a single value (total point value, TPV); showing the usefulness of the language resources in further processing; selection of these language resources that fulfilled predefined conditions. Considering the above mentioned criteria, the following PVs have been specified:

i. Availability
  i.1. Available for whom?
    company-internal (3)
    freely usable for PreR&D (2)
    freely usable for both PreR&D and R&D (1)
  i.2. At what price?
    Over 10K EUR (4)
    Between 1 and 10K EUR (3)
    Between 100 EUR and 1K EUR (2)
    Less than 100 EUR or free (1)
  i.3. How straightforward it is to reuse it (degree of adaptability)?
    Black box resource (3)
    Glass box resource (2)
    Open resource (1)
ii. Quality
  ii.1. Standard compliance (Is the resource based on a common standard?)
    No common standards used (3)
    No common standards used internally, but interfaces or converters to standards are available (2)
    Standard-compliant (1)
  ii.2. Soundness (Internal consistency, i.e., is the resource based on well-defined specifications?)
    No specifications available (3)
    Specifications cover only certain aspects of tools (2)
    Full specification exists (1)
  ii.3. Task-relevance (Is the resource suited for a specific task?)
    Not particularly well-suited, should be improved (3)
    To a certain extent (2)
    Very well suited (1)
  ii.4. Environment-relevance (Is the resource interoperable with other resources?)
    No (3)
    Yes, with a limited number of them (2)
    Yes, with many of them (1)

iii. Quantity (resources only)

Below 50 per cent of top quantity available for the language (3)
Between 50 and 90 per cent of top quantity (2)
Over 90 per cent of top quantity (1)

The lowest possible TPV is 8, the highest – 25. The established criteria for selecting language resources required TPV lower than or equal to the minimum value of 16. The TPV for resources being selected for the project could be calculated before any upgrade work. The process was directly related to lowering TPVs, which could and still can be used as a concrete indicator of project success.

Language White Papers

A certain measurement was made on LT resources and tools described in Language White Papers. In the Language White Paper series the respective LRs of the respective languages are described and ranked according to several criteria. The rating of the existing resources and tools was based on estimations by several leading experts using the following criteria (each ranging from 0 to 6).

i. Quantity: Does a tool/resource exist for the language at hand? The more tools/resources exist, the higher the rating.

0: no tools/resources whatsoever
6: many tools/resources, large variety

ii. Availability: Are tools/resources accessible, i.e., are they Open Source, freely usable on any platform or only available at a high price or under very restricted conditions?

0: practically all tools/resources are only available at a high price
6: a large amount of tools/resources is freely and openly available

iii. Quality: How well are the respective performance criteria of tools and quality indicators of resources met by the best available tools, applications or resources?

0: toy resource/tool
6: high-quality tool, human-quality annotations in a resource

iv. Coverage: To which degree do the best tools meet the respective coverage criteria? To which degree are resources representative of the target language or sublanguages?

0: special-purpose resource or tool, specific case, very small coverage, only to be used for very specific, non-general use cases

6: very broad coverage resource, very robust tool, widely applicable, many languages supported

v. Maturity: 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?

0: preliminary prototype, toy system, proof-of-concept, example resource exercise
6: immediately integratable / applicable component

vi. Sustainability: How well can the tool/resource be maintained/integrated into current IT systems?

0: completely proprietary, ad hoc data formats and APIs
6: full standard-compliance, fully documented

vii. Adaptability: How well can the best tools or resources be adapted/extended to new tasks/domains/genres/text types/use cases, etc.?

0: practically impossible to adapt a tool/resource to another task, impossible even with large amounts of resources or person months at hand

6: very high level of adaptability; adaptation also very easy and efficiently possible

Proportion between the selected resources developed inside and outside the consortium

The resources were also classified as being developed inside the consortium, outside the consortium or both. This information provided supplementary evidence that were compared with the gaps – thus, some efforts might be concentrated for further identification of language resources outside the consortium.

The combination of four indicators (each of them specified according to different sets of criteria) are used in the process of selection of CESAR language resources. The first indicator was general, thus assessing the indicator according to general yes/no criteria. All evaluated resources for a given language were listed within the criteria. All selected resources were and still are state-of-the-art representatives of their type. The above mentioned two indicators – Total point Value and Language White Papers –, are based on a numerical assessment of the resources according to previously established qualitative and quantitative criteria and conventions for their measurement. The preferable source of data for the analysis in CESAR are the tables for individual languages produced by the marks given for each of predefined categories in the Language White Papers. The fourth indicator is complementary – it is not of utmost importance for the selection itself but hints where the efforts should be put to fill the gaps in the selection.

The elaborated methodology is based on the combination of four indicators: general assessment, Total Point Value, Language White Papers and specification the origin of the resources.

The summaries presented in the deliverable *D2.4 Report on methodology and criteria for the selection of resources* highlighted on the data for language resources provided in the Language White papers. It shows that relatively equal results for all languages are visible in categories: Reference corpora, Parallel corpora, Speech corpora, Lexicon, Terminologies, and Thesauri, WordNets. On the other hand, the table also indicates the gaps of the resources for the target languages: Discourse Corpora, Semantics-Corpora, Multimedia and multimodal data and Ontological Resources for World Knowledge. It is not realistic to expect to fill the gap in the scope of the CESAR project but the clear understanding of necessities and clear definition of the future directions is of great importance.

The report illustrates how the adopted methodology and criteria were applied for each individual language: Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak. It gives an extensive overview and assessment of the selected language resources for the every language and identifies gaps in the provision of the language resources. For each language a profound analysis of the set of already selected resources and tools is performed. The analysis leads to the conclusion what kind of resources should be of the greatest interest in next rounds of selection.

Delivery of the resources and tools

After the careful selection of the language resources and tools a range of actions were provided in order to bring the selected resources into the envisaged form (cleared IPR, created an enhanced metadata description and the respective work on resources). The actions carried out on the selected reasources can be summarised as the following:

Metadata descriptions

The delivery or the resources and tools was preceded by a range of actions. One of them was the creation of the metadata schema, which was a shared task of all PSP project in META-NET Alliance. The metadata description scheme was prepared in tight cooperation of partner projects and reflects the needs of all CESAR partners. The description consists of a set of elements defined as recommended and facultative. The recommended set of metadata provided by META-NET forms the base of the work. However, the widely accepted document D7.2 of META-NET focuses on written corpora, within the CESAR project, a number of spoken and multimodal resources were also included. The document was used with expectations to cover all provided resources, but the detailed descriptions for some of their types were completed only at time of the second and third batch. META-NET has during the project extend its metadata description model to cover all other LRT types (spoken, multimodal, lexical resources and tools and technologies). Therefore the CESAR community had to pay a special attention to specify its metadata description model for non-textual LRTs. In the context of META-SHARE, the term metadata refers to descriptions of Language Resources and Tools, encompassing both data resources (textual, multimodal/multimedia and lexical data, grammars, language models etc.) and tools/technologies/services used for their processing.

The META-NET D7.2 document specifies a metadata model template, that was used for the description of Language Resources and Tools (LRTs) made available through META-SHARE (META-SHARE metadata model). The META-NET D7.2 puts the model in the context of its application (LRT sharing), delineating the intended goals of use and factors to be taken into account for its design. The proposed model was elaborated based on two main points: (i) user requirements, as collected through surveys and (ii) an overview of the most widespread metadata models and catalogue descriptions of LRTs (ELRA, LDC, and many others). Particular emphasis is put on the presentation of a minimal schema, which is a subset of the META-SHARE model, consisting of elements considered indispensable and hence compulsory for the description of LRTs.

Metadata descriptions of LRTs offered for META-SHARE are open distributed. Metadata descriptions follow a pre-defined, well described, machine readable format.

The harmonization of the description went with aims to agree on and apply standardized, optionally full-featured, flexible, machine readable metadata description requiring an obligatory minimal set of the most relevant resource/toll attributes. Detailed activity of the project partners during the project was as follows:

- Checking and adaptation of metadata schemes while providing to PSP partners a wide feedback;
- Working in close and continuous collaboration with other PSPs and META-NET, sharing of experiences, further emerging needs, problems, etc.;
- Providing and/or harmonizing and standardizing metadata descriptions for resources involved in the upload batches: included all mandatory as well as the most possible optional elements;
- Upgrading all metadata from previous batches to latest versions for next batches – as resources uploaded in the first two bathes might be adjusted while the project lifetime and the modifications were reflected also in the metadata schema.

For the final, 3rd upload batch (which also contained some adjusted resources from the previous batches), only some minor changes occurred in the metadata schemes. Already for the 2nd batch (July 2012) all metadata description schemes were available for all covered types (corpus, lexicon, language description, technology/tools) and all media (text, audio, video, image) of resources and tools. The used version of the metadata schemes for the final, 3rd batch was V3.0 which hat to be implemented in all META-SHARE nodes. The available V3.0. scheme set consists of a relatively large and complex set of XSD schemes supplied by META-NET and agreed between PSP partners. After META-SHARE nodes were set up, editing of the metadata could be done either on-line (through the metadata editor with built in validation methods) and off-line (using the same XSD schemes – but while the upload process the files are also validated with the built in mechanism.

Due to some modification of metadata schemes (for example, license notations were basically changed), batch 1 and 2 schemes had to be updated in M24. The conversion between versions 2.1 and 3.0 was controlled by the contributing Partners, including hand-made checking and validation of all XML metadata description files. Revised files were either reimported into the META-SHARE managing node or imported into the Partners' network node for harvesting by the managing node.

The minimal schema

Partners agreed in providing the metadata description covering at least the minimal schema by the 31th January 2013. However, the description should be as complete as possible and cover possibly non mandatory elements as well in order to provide more detailed information on the resources involved.

CESAR partners have reported several problems to META-SHARE metadata editor software developers during the first upload batch and related to version 1. Version V2.1 and also current version V3.1 (current, at the time of compilation of the report) of the software were experienced more stable and reliable. This latter approach allows for some flexibility, however, validation and re-checking during the import is crucial as offline editing is more prone to errors.

META-SHARE nodes

Official CESAR nodes for META-SHARE

For batches 1 and 2, Partners agreed to set up one official META-SHARE node in Warsaw, Poland, maintained by IPIPAN. The same node was used for batch 2.

At the end of the project, META-SHARE nodes were organized into a hierarchical structure: managing nodes are synchronized, and provide all META-SHARE metadata and resources, whilst network nodes are not synchronized, but harvested by a managing node.

As the META-SHARE server software runs in full functionality (including synchronization), CESAR Partners decided to promote the original Warsaw node to become a managing node (CESAR managing node), and set up a network node at each Partner's premises to provide metadata for harvesting by the CESAR managing node, which shares metadata with other META-SHARE managing nodes.

Some of the network nodes operate only from early February 2013 as the recent bug-fix version of META-SHARE software was released at the end of January 2013, and installation needed some more time.

Sustainability

Partners and especially IPIPAN express their wish and commitment to maintain and run the META-SHARE managing node for CESAR after the project ends, at least for a period of two years. This commitment involves all resources referenced in the META-SHARE nodes, but hosted physically elsewhere (according to the letters of intent the Partners submitted for META-FORUM 2012 in Brussels, 2012).

The official META-SHARE managing node for CESAR is available at:
http://nlp.ipipan.waw.pl/metashare

All CESAR Partners have received user accounts and passwords to be able to edit their metadata. The server was set up end of October, 2011. Update for version 2.1 was carried out in June, 2012, for version 3.1 in January, 2013.

CESAR network nodes already in service are:

- RIL MTA node: http://metashare.nytud.hu/,
- BME-TMIT node: http://metashare.tmit.bme.hu/,
- FFZG node: http://meta-share.ffzg.hr/,
- IPIPAN node: http://nlp.ipipan.waw.pl/metashare/,
- ULodz node:http://metashare.ia.uni.lodz.pl/,
- IBL node: http://metashare.ibl.bas.bg/,
- UBG node: http://meta-share.matf.bg.ac.rs
- LSIL node: https://metashare.korpus.sk/

Deliverable of the chosen language resources and tools (provided and uploaded in three batches)

The selected resources were delivered in three batches:

- 1st batch: 1 December, 2011
- 2nd batch: 1 August, 2012
- 3rd batch: 1 February, 2013

The enhancement of the tools and resources were carried out with three types of activity (as agreed within META-NET): resource upgrade, extension and cross-lingual alignment. The initial set of language resources was during the project extended from 132 to 251 pieces of LR, labeled with CESAR.

Upgrading resources

In resources where it was necessary, the consortium (particular partners) made upgrade on the selected resources to standards agreed on jointly with other projects and META-NET, in particular the META-SHARE initiative and partners. In particular, the partners improved the documentation of the resources in question (both formal - metadata – and informal - narrative), removed certain bugs or inconsistencies and cleaned the datasets. The main idea of the approach was to make the resources compliant with (or mappable onto) widely recognized technical and linguistic standards (e.g. character sets, tag sets). For the description of resources and in order to make their harvesting possible, standards and protocols provided by META-SHARE were complied with.

For collecting and re-using repository data, the consortium followed the META-SHARE approach and adopt metadata harvesting using the Open Access Initiative Protocol for Metadata Harvesting (OAI-PMH). The consortium likewise adopted widely accepted standards promoted by META-

SHARE such as Unicode (ISO 10646) for text encoding, ISO 639 for language codes, XML for content and metadata representation.

The upgrade task was focused on reaching META-SHARE compliance. During the task of upgrade, the following activities were done on the following LRs:

- upgrade for interoperability (changing annotation format, type, tagset),
  - Polish Sejm Corpus: conversion to TEI P5 and NKJP tagset,
  - Polish WordNet: conversion to WordNet-LMF,
  - Polish Named Entity Resources: extraction of the open source resources from the mixed open and proprietary data (removing copyrighted inhabitant names and relational adjectives stemming from Polish settlements),
  - LUNA.PL and LUNA-WOZ.PL: conversion to TEI P5,
  - Conversion of the PELCRA Conversational Corpus to TEI P5,
  - Conversion of PELCRA parallel corpora to XLiFF and TEI P5,
  - SrpLemCor: a part of SrpCor made available as open source (clearing the copyright issues),
  - SrpFranKor: all texts in basic TEI P5, all bi-texts in TMX,
  - Verne80days: all texts in basic TEI Pt, all pairs in TMX,
  - Bulgarian Sense□annotated Corpus: Annotation upgrade,
  - Bulgarian X language Parallel Corpus: Linguistic preprocessing, annotation and alignment.
  - PNEG, HVPC: conversion to and LMF-compliant format, validation.
  - TaCo: standardization of representation of linguistic information: tagset definition based on Spejd formalism, XCES as input-output format.
  - N□grams from balanced NKJP: extraction of plain-text content from the corpus and converting all characters to lower-case, extraction of all required n-grams from the above-mentioned content, sorting the result by the number of unique occurrences.
  - Redistributable subcorpus of the National Corpus of Polish: extraction of freely distributable texts from NKJP.
  - SEJF, SEJFEK, SAWA, PNET: proofreading of the lexicon and generation of its extensional form (containing all inflectional and syntactic variants).
  - Walenty: automatic conversion of entries from the electronic version of Świdziński's valence dictionary to the established format of the new valency dictionary.
  - Polish Wikipedia corpus: creation of a new format for the resource.
  - Szeged□ Treebank□FX: the annotation was mapped to the dependency version of the treebank
  - HUCOMTECH□ multimodal□ database: Conversion of the annotations to ELAN (.eaf) format
  - BEA□ Hungarian□ spontaneous □speech □database: anonymisation of the sound files and transcription
  - Hungarian □kindergarten □language□ corpus: anonymisation of the sound files and transcription using CHAT format of CHILDES
  - ht□online: conversion of the database to the common formats
  - Hungarian □concise □dictionary□ (with□ sample□ sentences): XML conversion to TEI P5
  - N□grams □from □Hungarian□ National □Corpus: extraction of plain-text from HNC and converting to lover-case charcters, generating n-grams
  - Hungarian MALACH Speech Database: standardized speech and annotation
  - Hungarian Medical Speech Database: standardized speech and annotation formats
  - Conversion of the PELCRA Conversational Corpus to TEI P5,
  - Conversion of PELCRA parallel corpora to XLiFF and TEI P5,
  - Conversion of PELCRA Learner corpus to TEI P5
  - n-grams from Croatian National Corpus: existing procedures have been upgraded to be compatible with the methodology agreed upon by CESAR partners
  - Croatian Translations of Acquis Communautaire: JRC DTD validation
  - Orwell 1984 Croatian: compliant with MulTextEast v4.0
  - Croatian Wordnet: conversion from existing format into one usable with other editors (Hydra, Wordnetloom)
  - Slovak National Corpus: individual document metadata have been converted to TEI P5
  - n-grams from Slovak National Corpus: existing procedures have been upgraded to be compatible with the methodology agreed upon by CESAR partners; a new set of n-grams has

been released, based on the new version of the Slovak corpora; the order has been increased to 4

- SrpLemCor: a part of SrpCor made available as open source (clearing the copyright issues),
- SrpFranKor: all texts in basic TEI P5, all bi-texts in TMX,
- SrpEngKor: all texts in basic TEI P5, all bi-texts in TMX
- Verne80days: all texts in basic TEI P5, all pairs in TMX,
- SrpMD: Serbian Morphological Dictionary converted from NooJ format to Multext-East;
- Verne80DaysMSD: Serbian translation of Verne’s novel “Around the World in 80Days” morphosyntactically tagged and disambiguated converted from Nooj to Multext-East format;
- SrpNEval: Named Entities evaluation corpus for Serbian compiled from various texts automatically tagged with NE and manually corrected,
- SrpNGrams: set of N-grams extracted from Serbian Lemmatized and PoS Annotated Corpus (SrpLemKor) for N from 1 to 5. Each unigram is maximum continuous chunk of non-whitespace lower-case characters; the methodology agreed upon by CESAR partners
- Bulgarian Sense□annotated Corpus: annotation update.
- Bulgarian X language Parallel Corpus: Linguistic preprocessing, annotation and alignment;
- Bulgarian wordnet: upgrade to compatibility with Princeton wordnet 3.0.

- technology-related upgrade (wrapping, refactoring, etc.),
  - NERF: upgrade of the model for NE recognition,
  - Hungarian WordNet: The database has been filtered for BCS concepts,
  - Mindentudás□ Speech □Corpus and Hungarian Broadcast News Database: extract audio tracks from video,
  - Hungarian Word □Level □Speech□ Database: manual marking of sound boundaries and sound symbols on the waveform,
  - Slovak National Corpus: web interface access added (conversion to NoSketch Engine),
  - Organizing Digitized Material: The first version of this software tool was produced for organizing digitized cultural heritage material belonging to ethnographic maps of Serbia,
  - Wordnet web service: Development of the wordnet database, development of the web service,
  - Bulgarian Spell Checker for Windows, Bulgarian Spell Checker Web Service: Implementation of the Spell Checker engine, Development of spell checker dictionary.
  - NERF: reimplementation in Haskell,
  - Hungarian National Corpus: new types of analysis were made
  - Hungarian□NER□Corpus□based□on□Wikipedia: technology related upgrade (download, parsing and cleaning of the XML-files, NE-labeling), enhancement of the NE-tagger
  - Hungarian□Opinion□Tagged□Sentence□Bank: upgrade of the NER tools
  - Hungarian Phonetic Transcriber: enhance phoneme set and transcription rules
  - Automatic Prosodic Segmenter: retrain prosodic models on checked transcripts
  - Graphical query interface for Hungarian Read Speech Precisely Labelled Parallel Speech Corpus Collection: development of the web service
  - Organizing Digitized Material: The first version of this software tool was produced for organizing digitized cultural heritage material belonging to ethnographic maps of Serbia,
  - eEmotion: a web application for ontological based emotions recognition and tagging of Serbian texts,
  - Wordnet web service: Development of the wordnet database, development of the web service,
  - Bulgarian Spell Checker for Windows, Bulgarian Spell Checker Web Service: Implementation of the Spell Checker engine, Development of spell checker dictionary.
  - Hydra and Chooser: code refactoring; Hydra: simplifying table descriptors.
  - Speech Analyser Rapid Plot (SARP): upgrade of the tool.
  - Translation Reference Library (TREFL): upgrade of the tool.
  - Real Time Comparison (RTComp): devlopment and upgrade of the tool.

- application of techniques of finding inconsistencies and errors in (automatically and/or manually built) linguistic resources, incl. corpora and lexica,
  - Szeged Corpus 2.0 and Szeged Treebank 2.0: Annotation errors have been corrected (concerning lemmas and POS-tags), Errors in the XML structure have been corrected,
  - Szeged Named Entity Recognition Corpus: Annotation errors have been corrected,

  - Hungarian WordNet: annotation errors have been corrected concerning definitions of synsets, format and non-lexicalized synsets,
  - Mindentudás□Speech□Corpus: check for and correction of inconsistencies
  - Hungarian□ Speech □Emotion□ Database: anonymization, error corrections (automatic+manual check), filter inconsistencies,
  - Hungarian Word □Level □Speech□ Database: Checking and optimizing the waveforms of the word items (silent part- word – silent part), unification of the silent periods at the beginning and ending part of the wave file, amplitude correction on word waveforms, manual checking of the marked sound boundaries and the given sound symbols,
  - Croatian-English Parallel Corpus has been checked for alignment,
  - Croatian translation of Orwell 1984 has been MSD-tagged and lemmatised in conformant with MulTextEast resources v4.0
  - French□Serbian Aligned Corpus: were corrected and new bi-texts in TMX format were produced,
  - Polish Sejm Corpus: semi-automatic correction of some common typos,
  - NKJP: using CorpCor tool for error detection and manual correction of samples,
  - Hungarian□WSD□Corpus, Szeged□Criminal□NE□Corpus: XML errors have been corrected
  - Szeged□Criminal□NE□Corpus: annotation errors have been corrected
  - Slovak National Corpus: automatic correction of several types conversion errors
  - Named entity lexical database: check items
  - Hungarian formant database: check items
  - Hungarian Medical Speech Database: remove lower SNR recordings
  - Hungarian MALACH Speech Database: validation of transcripts
  - Hungarian Broadcast Conversation Database from the Catholic Radio of Eger (EKR): check records
  - Croatian Morphological Lexicon v5.0: manual checking of new lemmas
  - Croatian Wordnet: manual checking of synsets
  - Orwell 1984 Croatian: manual checking of MSD-tagging/lemmatisation
  - South-East European Parallel Corpus: correcting encoding errors
  - Croatian Dependency Treebank: manual checking of dependency relations
  - Manutally Annotated Slovak Corpus: semi-automatic correction of sentence segmentation, automatic error detection in morphological tagging and desambiguation
  - Parallel Slovak corpora (Slovak-English and Slovak-Czech): automatic alignment verification
  - SrpWN: fixing hanging links and duplicate literals,
  - SrpFranKor: links in all bi-texts checked and corrected,
  - SrpEngKor: links in all bi-texts checked and corrected,
  - Verne80days: links in all pairs checked and corrected,
  - Verne80daysMSD: errors in manual disambiguation corrected,
  - SrpNEval: errors in manual evaluation corrected.
- metadata-related work (creation, enchancement, conversion, standarization),
  - all resources: descriptions harmonized with META-SHARE metadata model.
- harmonization of documentation (conversion to open formats, reformating, linking),
  - LUNA.PL and LUNA-WOZ.PL: documentation of the TEI P5 annotation, linking to documentation of another version of the corpus with MLF transcription,
  - Szeged Corpus 2.0 and Szeged Treebank 2.0: update and English translation of the documentation,
  - Hungarian WordNet: documentation of the corpus has been updated, expanded and translated to English,
  - Hungarian Webcorpus: improving the documentation for the downloadable, and writing new documentation for the web-based frequency dictionary,
  - Hunglish Parallel Corpus Version 2.0: normalization of the documents for format and encoding with a combination of manual and automatic processing,
  - morphdb.hu, hunmorph, huntoken: improvements to the documentation,
  - Mindentudás□ Speech□ Corpus: improve and extend documentation,
  - Hungarian□ Broadcast□ News □Database: documentation harmonization: reformatting,
  - Sound □Gesture □Database and Hungarian□ Speech □Emotion□ Database: newly create documentation,
  - Croatian□English Parallel Corpus: several types of conversion,
  - Croatian Vallency Dictionary: documentation on logical structure of CROVALLEX,

  - Croatian Wordnet: conversion from existing format into one usable with other editors (Hydra, Wordnetloom),
  - Bulgarian National Corpus Collocation service, Bulgarian Part□of–Speech Corpus: conversion of the corpus format.
  - Hungarian Medical Speech Database: provide documentation
  - Automatic Prosodic Segmenter: extend documentation
  - Hungarian Phonetic Transcriber: provide user manual
  - Spejd, SEJFEK4Spejd: standardization of documentation,
  - Slowal, PoliMorf/Lexeme Forge: preparation of user manuals.
  - Slovak National Corpus: comprehensive user manual, tutorial and documentation has been written
  - Croatian□English Parallel Corpus: several types of conversion,
  - Croatian Vallency Dictionary: documentation on logical structure of CROVALLEX,
  - Croatian Wordnet: conversion from existing format into one usable with other editors (Hydra, Wordnetloom),
  - Croatian National Corpus v3.0: new documentation on the corpus web site
  - Bulgarian National Corpus Collocation service, Bulgarian Part□of–Speech Corpus: conversion of the corpus format.
  - Hydra and Chooser: preparation of installation and user manuals and making them available.
  - SrpNooj: Serbian Nooj modul was produced consisting of Serbian morphosyntactic dictionaries, example text, dictionary properties' definition file, example morphological and syntactic grammars (in two scripts)

- preparation for maintenance and deployment (debugging, cleaning, building test environments, preparing code repositories),
  - LUNA.PL and LUNA-WOZ.PL: preparation of the ODD files used to create RNG schemata,
  - Szeged Corpus 2.0 and Szeged Treebank 2.0: revision of the MSD coding system,
  - Hungarian Word □Level □Speech□ Database: software development for generating unified acoustic images from the waveforms, the sound boundary markers and sound symbols.
  - Morfeusz, morfologik-stemming: using newest morphological data and inflection patterns exported from PoliMorf.
  - Morfologik: development of a new web tool to maintain the dictionary.
  - Walenty: creating a web tool allowing manual edition of the valence frames.
  - Pantera: redesign of the library API.
  - NERF: implementation divided into a collection of packages which can be developed and improved independently.
  - Prolexbase: a tool has been developed in order to populate Prolexbase from open data.
  - CollTerm: parametrisation of the tool provided with parameter files.
  - Web Content Extractor: tool preparation for publishing, code cleanup and optimisation.
  - Corpus Aligner: adjusting I/O format to TMX standard, debugging.
  - Croatian National Corpus: redesigning corpora interface (migration to Bonito2 browser client)
  - Slovak National Corpus: redesigning corpora interfaces environment (both monolingual and parallel), cluster based deployment (enhances availability, redundancy and long term support)
  - SrpKor, SrpFraKor, SrpEngKor: redesigning corpora interfaces environment (both monolingual and parallel); database approach applied for maintenance of texts and their meta-data,
  - SrpSpell: Serbian Spell Checker web service,
  - Automatic Prosodic Segmenter: cleaning
  - Bulgarian Spell Checker web service: error fixing.
  - Bulgarian Spell Checker for Windows, Bulgarian Spell Checker for Mac OS, Bulgarian Spell Checker web service: debugging; errors and ambiguities resolved.
  - Hydra and Chooser: debugging.
  - Bulgarian Sentence Splitter and Tokeniser: debugging.
  - Bulgarian wordnet: verification of consistency of the data

- programming tasks (bug-fixing and standardizing API calls).
  - NERF: bug-fixing,
  - SrpWN: fixing hanging links,
  - SrpFranKor: links in all bi-texts checked and corrected,

    - Verne80days: links in all pairs checked and corrected,
    - Szeged Corpus 2.0 and Szeged Treebank 2.0: new editor has been developed for correcting misspelled words,
    - Hungarian Webcorpus: bugfixes and minor enhancements to the web-based interface,
    - Hunglish Parallel Corpus Version 2.0: fine-tuned and debugged the hunalign-harness pipeline,
    - hunalign: design of a new parallel text processing pipeline,
    - Hungarian□ Speech □Emotion□ Database: develop automatic anonymizer (post manual check) and structure editor tool,
    - Croatian Lemmatisation server: enhanced version of HML inserted into server, starting to work on turning web interface into a proper web service,
    - Pantera: improvements in sentence segmenter.
    - NERF: added support for external dictionaries.
    - LexemeForge: customizable dictionary exports have been implemented.
    - Croatian Web Services: programming and connecting the modules, standardisation of I/O protocols (REST)
    - Slowal: implementation of new functionalities.
    - Multiservice: introducing the Apache Thrift based API used to plug-in new language tool.
    - Development of programmatic interfaces to the PELCRA HASK collocation dictionaries
    - hunner, hunpars, hunpos: bugfixing and other programming developments
    - Graphical query interface for Hungarian Read Speech Precisely Labelled Parallel Speech Corpus Collection: development, GUI programming
    - Bibliša: developemnt of a web application for search of digital libraries of articles from bilingual e-journals,
    - NERanka: development of a web application for automatic NE tagging of Serbian texts

- IPR issues:
    - Hungarian WordNet: Negotiation on licensing issues.
    - Mindentudás□ Speech□ Corpus: extensive negotiation on licencing, involving intensive communication with META-NET and ELRA,
    - Hungarian□ Broadcast □News □Database and Hungarian□ Speech □Emotion□ Database: demand for MS-NoRedistribution licences fulfilled,
    - Serbian Lemmatized and PoS tagged corpus,
    - SEJF, SEJFEK, SAWA: the resource made available under the 2-clause BSD licence (FreeBSD).
    - Summarizer: the resource made available under CC-BY licence.
    - WikiTopoPl: the resource made available under the CC-BY-SA 3.0 Unported license.
    - PolNet 1.1: formal clarification of the IPR status.
    - Acquisition and release of SNUV speech database under CC-BY
    - POLFIE: available under the GPL (version 3) license.
    - CorpCor: the code has been released under the GPL v. 3 license.
    - Syntactic-Generative Dictionary of Polish Verbs: released under CC-BY after several years of struggle.
    - HuWN: IPR issues clarified
    - Croatian resources published under respective licences through META-SHARE
    - Authors' promise to release the dictionary database of "Dictionary of Slovak Collocations. Adjectives" and "Dictionary of Slovak Collocations. Nouns." under CC-BY-SA after the printed version is published has been obtained
    - Graphical query interface for Hungarian Read Speech Precisely Labelled Parallel Speech Corpus Collection: IPR discussion with IPR holder for META-SHARE deposition
    - Hungarian Medical Speech Database: Consortium Agreement between IPR-holders for META-SHARE deposition
    - Hungarian MALACH Speech Database: IPR discussion with IPR holder for META-SHARE deposition
    - Hungarian Broadcast Conversation Database from the Catholic Radio of Eger (EKR): IPR discussion with IPR holder for META-SHARE deposition
    - SerLemKor: Serbian Lemmatized and PoS tagged corpus.
    - SrpNovKor: Corpus of Contemporary Serbian Newpapers and Magazines made available for commercial use by IPR holder;
    - SrpRetFig: Database of Rhetorical Figures for Serbian made freely available for non-commercial use by IPR holder.

- Hydra and Chooser: available under GPLv3 license.
- Bulgarian National Corpus (BulNC), Bulgarian-X language Parallel Corpus (Bul-X-Cor), Bulgarian Part-of–Speech Corpus (BulPosCor), Bulgarian Sense-Annotated Corpus (BulSemCor): results of corpora are accessible under META-SHARE NoRedistribution Non-Commercial license.
- Bulgarian Spell Checker for Windows, Bulgarian Spell Checker for Mac OS, Bulgarian Spell Checker Web Service, Bulgarian Sentence Splitter and Tokenizer: available under META-SHARE NoRedistribution Non-Commercial license.
- Lists of Bulgarian Multiword Expressions, Bulgarian MWE dictionary, BgMWE, Bulgarian Frequency Dictionary, N-grams from Bulgarian National Corpus (BgNgrams): available under META-SHARE NoRedistribution Non-Commercial license.
- Bulgarian Grammar checker, Web based infrastructure for Bulgarian data processing: results available under META-SHARE NoRedistribution Non-Commercial license.
- Mutilingual dictionaries: available under META-SHARE NoRedistribution Non-Commercial license.
- TextMatch, Bulgarian Automatic Collocations Dictionary: results available under META-SHARE NoRedistribution Non-Commercial license.
- Dictionary of Synonyms in Bulgarian Language, Dictionary of Antonyms in Bulgarian Language, Register of Phraseologisms in Bulgarian Language, Dictionary of Neologisms in Bulgarian Language, Bibliography of Bulgarian Lexicology, Phraseology and Lexicography: available under META-SHARE NoRedistribution Non-Commercial license.

The upgrade of tools was made with aim to promote language-independent tools among consortium partners where available and/or applicable, e.g. to enhance resources by adding new levels of annotation.

The selection of resources for upgrade was carried out with the following principles in mind:

- the resources are state-of-the-art representatives of their type for a certain language,
- if more than one valuable representative of certain tool type for a language was available (e.g.
- two morphosyntactic analysers with equally popular tagsets or formal grammars used for
- different purposes), all of them are included in the selection,
- licencing issues allow to freely process and make available the resources and resource-related materials or the consortium succeeds in reaching an agreement with respective copyright holders.

Extending and linking resources

Selected existing resources were extended or interlinked to improve their usability and provide the processing toolkit containing standardized set of basic language tools for all project languages. To increase coverage, the actions were synchronized with the outcome of charting the national scene activity, which identified language gaps related to either inefficient technology or insufficient activity.

The list of actions done is strongly related to resource types and includes: population with additional data (for corpora, WordNets), improving levels of linguistic analysis (corpora, dictionaries) or general extension (e.g. formal grammars). Existing corpora along with information extraction tools were used to extend corpus-powered tools such as valency dictionaries or WordNets.

Monolingual resources were gathered first to cross-link, align them etc., e.g. for the purpose of building statistical models to be used for Machine Translation.

The chosen resources were extended or linked across different sources to improve their coverage and increase their suitability for both research and development work. This task took into account the specific goals of the project, identified gaps in the respective language community, and most relevant application domains.

Selection of resources extended/linked was based on those made available within task 3.1 to further enhance a smaller, but well-defined set of resources. Following rationale was applied:

- 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 (e.g. extended grammars to existing shallow parsers) rather than major restructuring,

- additional resources were integrated with existing ones only if they significantly improved the quality of resulting resources,
- if more than one representative of certain tool type for a language has been selected in task 3.1, they were very likely to be interlinked to benefit from strong points of both solutions (unless their usage patterns do not encourage such action),
- if less-developed, but still very popular (at least within one language community) tools cuold benefit from the enhancement basing on their well-developed equivalent (provided that no extensive work was necessary and that the latter tool cannot be used as a building block in further applications of the former tool), their enhancement is also considered,
- tools offering language-neutrality or cross-linguality are preferred.

The extension/linking task consisted of following actions on the following LRs:

- adding new portions of data, enhancement of resources, interlinking resources:
  - Polish Sejm Corpus: adding transcripts from official parliamentary questions/answers,
  - PoliMorf: new portions of manually verified data (after performing the merger of SGJP and Morfologik dictionaries),
  - plWordNet: new portion of data created by semi-automatic extension of the previous version,
  - ProlexBase: 165,000 inflected forms for Polish names have been automatically generated and manually validated.
  - Corpus of the Polish language of the 1960s: manual annotation of the corpus texts (segmentation and morphosyntactic level).
  - Hungarian National Corpus: was extended up to 1200 million words
  - Hungarian Language □Processing□ Tools □in □NooJ: upgrade of the dictionaries
  - New version of Slovak National Corpus and related subcorpora has been released, the size of the main corpus reached 1200 million words
  - New version of Corpus of Spoken Slovak reached 2.6 million words
  - Slovak Morphology database has been increased to 97 thousand lemmata
  - Slovak-Czech and Slovak English parallel corpora were increased by including more texts and also by adding separate corpora of freely downloadable texts; the final sizes are 6.4 and 10 millllion sentence pairs, respectively
  - Slovak Terminology database has been extended by several hundered terms; a new field (Computer Science) has been added
  - SrpWN: extended from 15,200 synsets to 18,366 synsets and adding new relations,
  - SrpKor: enhancement of the corpus available on the Web – from 23 million words to more than 113 million words (several technical enhancements made); corpus is lemmatized and PoS tagged; all text classified using UDC; the user interface was significantly improved.
  - Verne80days: addition of two new languages: Albanian, Slovenian and Hungarian, and new language pairs
  - SrpFraKor: enhancement of French-Serbian Aligned Corpus with new aligned texts from various domains: literature, newspaper and scientific texts.
  - SrpEngKor: enhancement of English-Serbian Aligned Corpus with new aligned texts from various domains: literature, newspaper and scientific texts.
  - Hungarian □BABEL: phoneme segmentation, syntactic annotation, linked prosodic pre-processing,
  - Croatian Wordnet: enlargement with new synsets, literals and relations,
  - Croatian National Corpus v3.0: extended to 231% of its v2.5 size with different new texts added, MSD-tagging and lemmatisation.
  - Corpus of Narodne novine: recrawled years 1990-2004, crawled years 2005-2012, MSD-tagging and lemmatisation.
  - Croatian Web Corpus: new MSD-tagging/lemmatisation
  - Slovene Web Corpus: new MSD-tagging/lemmatisation
  - South-East European Parallel Corpus: recrawled
  - Croatian-English WebParallel Corpus: crawling, conversion
  - Croatian Morphological Lexicon v5.0 extended with additional 12,000 lemmas
  - CESAR Aligned Wikipedia Headword list: collecting headwords
  - Croatian Translations of Acquis: collection, conversion
  - Orwell 1984 Croatian: MSD-tagging and lemmatisation.
  - Croatian Sentiment Lexicon: enlargement of the lexicon.
  - Hungarian Medical Speech Database: create database from scratch

  - Hungarian Broadcast Conversation Database from the Catholic Radio of Eger (EKR): extension
  - Bulgarian National Corpus: increasing the size of the corpus in a balanced way (up to over 1.2 billion words).
  - Wiki1000+: development of the corpus and its integration into the BulNC.
  - Bulgarian X-Language Parallel Corpus; increasing the size of the corpus (up to over 5.4 billion words, including the Bulgarian core; with texts in languages different than Bulgarian up to 4.2 billion words).
  - Bulgarian Sentence- and Clause-Aligned Corpus: development of the corpus and its integration into the Bulgarian X-language Parallel Corpus.
  - Bulgarian wordnet: Enlargement of Bulgarian wordnet with new synsets, literals and relations (up to over 49,000 synsets).
  - Corpus of Spoken Bulgarian: extended to 523,128 signs.
  - Corpus of Colloquial Bulgarian: extended to 357,584 signs.
- interlinking resources:
  - PoliMorf, resulting from merging of Morfeusz SGJP and Morfologik,
- linking existing resources across different sources,
  - plWikiEcono: linking a corpus of Polish Wikipedia articles from the domain of economy with NKJP,
  - Bibliša: a web application for search of digital libraries of articles from bilingual e-journals links various multilingual resources (including Serbian): Wordnets, terminological databases, morphosyntactic dictionaries,
  - NERanka: a web application for automatic NE tagging of Serbian texts links various resources for Serbian: conversion of scripts, morphosyntactic dictionaries, local grammars and syntactic grammars for NER,
  -
- providing building blocks to the existing tools (e.g. extended grammars to existing shallow parsers)
  - Prosodic Segmenter: provide a prosodic model for a widespread open-source speech recognition tool (HTK),
  - Croatian and Slovene NERC models for Stanford NERC: training texts manually annotated and used for training
- major restructuring,
  - Hunglish Parallel Corpus Version 2.0: redesigned filename conventions and directory organization,
  - Hungarian□ Speech □Emotion□ Database and Sound□ Gesture □Database: completely new structure, splitting long speech files into separate ones, new naming conventions,
  - Mindentudás □Speech □Corpus: create and provide audio tracks,
  - Hungarian Medical Speech Database: create database from raw recordings, organize structure.
- integration of additional resources with existing ones to improve the quality of resulting resources,
  - hunalign: integration of new tools in order to improve the final quality,
  - Hungarian□ Speech □Emotion□ Database and Hungarian Telephone Client Speech Database (2nd batch): integration, merging wherever possible (in order to obtain an interlinked verbal+non-verbal corpus),
  - Hungarian□ Speech □Emotion□ Database and Hungarian Telephone Client Speech Database: integration, merging wherever possible (in order to obtain an interlinked verbal+non-verbal corpus).
- enhancement of resources.
  - SrpWN: extended with 1,691 synsets and adding new relations,
  - SrpKor: enhancement of the corpus available on the Web; corpus is lemmatized and PoS tagged; all text classified using UDC,
  - Verne80days: addition of two new languages: Albanian and Hungarian,
  - Hungarian □BABEL: phoneme segmentation, syntactic annotation, linked prosodic pre-processing (enhancements will be released with 2nd batch),
  - Sound □Gesture □Database (Hungarian): extended by 20% (new gestures added),
  - Slovak National Corpus: extended by 36% (many different texts added),
  - Corpus of Spoken Slovak: extended by 140% (recordings added to cover the territory of Slovakia more uniformly),

    - Croatian National Corpus extended by 30% with different new texts added,
    - Croatian Morphological Lexicon extended with additional 12,000 lemmas,
    - Croatian Wordnet: enlargement with new synsets, literals and relations,
    - Croatian-English Parallel Corpus has been extended with additional 80,000 aligned sentences (ca 103% increase) crawled from web,
    - Serbian Wordnet: enhanced, from 15,200 synsets to 16,891 synsets,
    - Corpus of Contemporary Serbian: The part of corpus available on web was enhanced – from 23 million words to more than 113 million words (several technical enhancements made),
    - Bulgarian National Corpus: Increasing the size of the corpus in a balanced way,
    - Bulgarian wordnet: Enlargement of Bulgarian wordnet with new synsets, literals and relations.

Aligning resources across languages

Relatively smaller number of tools and resources were used to propagate their formats, processing methods and technologies to their foreign equivalents. Naturally, parallel corpora were used as basis for cross-lingual alignment for other project languages, thereby creating thus new useful resources.

Tools with regional reach were extended with modules for other project languages and equipped with cross-lingual annotation mechanisms.

The cross-lingual alignment consisted of the following actions on the following LRs:

- introducing language-neutrality,
    - Prolexbase: 65,500 language-independent relations have been extracted and manually validated.
    - Hungarian□ Historical Corpus: genre alignment with Hungarian National Corpus
- introducing cross-linguality,
    - automatic alignment of the Polish CORDIS and RAPID parallel corpora
    - manual alignment of the Academia parallel corpus
    - development of cross-language alignment methods for Polish and English dictionaries off collocations
    - Multilingual Edition of Verne's Novel "Around the World in 80 Days": Two new languages were added for this release: Hungarian and Albanian,
    - automatic alignment of the enhancement of Croatian-English parallel corpus with sentences crawled from web,
    - plWordNet: adding alignment with Princeton WordNet,
    - SrpWN: adding alignment with Princeton WordNet 3.0,
    - multilingual lexicon of toponyms: alignment of a resource coming from a different project with CESAR languages,
    - Prolexbase: 40,000 Polish, 33,000 English and 20,000 new French proper names have been extracted from Wikipedia and GeoNames, interlinked, manually validated and inserted in Prolexbase.
    - OpenCyc: 13,000 symbols translated into Polish.
    - SzegedParallel, SzegedParallelFX: corrections in alignment errors
    - automatic alignment of the Polish CORDIS and RAPID parallel corpora
    - manual alignment of the Academia parallel corpus
    - development of cross-language alignment methods for Polish and English dictionaries off collocations
    - alignement of Bulgarian, Croatian, Hungarian, Serbian and Slovak wordnets based on Princenton WordNet mappings resulted in the "Multilingual Glossary of Synsets" resource
    - Multilingual Edition of Verne's Novel "Around the World in 80 Days": Two new languages were added for this release: Hungarian and Albanian,
    - SrpRetFig: a database of Serbian rhetorical figures related to rhetorical figures for English,
    - automatic alignment of the enhancement of Croatian-English parallel corpus with sentences crawled from web
    - Croatian Wordnet: alignment with Princeton WordNet 3.0
    - CESAR Aligned Wikipedia Headword list: aligning headwords
    - Croatian Translations of Acquis: alignment (TMX)
    - Croatian-English Web Parallel Corpus: alignment (TMX)
    - Bulgarian X-language Parallel Corpus: automatic alignment of portions of the corpus.
    - Bulgarian-English Sentence- and Claused-Aligned Corpus: automatic alignment and manual verification.

- mapping between tagsets
  - NERosette: a web application for retrieval of aligned texts; enables mapping of various NE tagging schemes to a chosen one,

- mapping between outputs and inputs of linguistic tools for particular language,
- synchronization of resources available for consortium languages,
  - Automatic Collocation Dictionaries produced for all project languages on the basis of new Sketch Grammars developed for some CESAR languages,
  - NERosette: a web application for retrieval of aligned texts synchronizes NE tagged aligned texts,
  - CESAR Aligned Wikipedia Headword list (incl. English)
  - Multilingual Glossary of Synsets (incl. English)

- extension of language models to embrace cross-linguality and/or promote language independence.

Cross-lingual alignment of resources, as the most demanding task, was applied only to a small number of resources. During the task, the following approaches were followed:

- application of techniques of mapping between tagsets and, more generally, outputs and inputs of linguistic tools for particular language,
- synchronization of resources available for consortium languages,
- extension of language models to embrace cross-linguality and/or promote language-independence.

During the aligning of selected resources the following rationale was applied:

- no more than a tool of a certain type for each language was used in the process,
- whenever applicable, the largest set of languages was selected (preferably with English as a hub language; the languages going beyond natural consortium scope of interest are not excluded),
- language-independence was targeted to a great extent.

During the CESAR-project resources were delivered in the following sets:

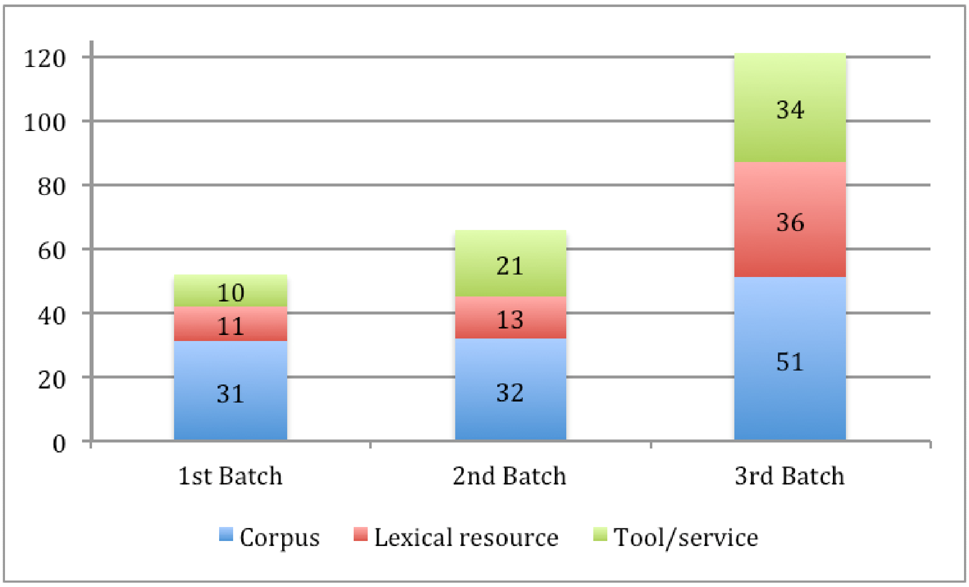


Figure 1. CESAR resources in numbers

Special mention must be made of work to upgrade, extend and cross-link of the finite state tool NooJ (www.nooj4nlp.net).

The progress of the development in NooJ was made mostly by IPUP. Development of NooJ was aligned with the following activities as written in the DoW:

- Make NooJ open source
- Make NooJ platform independent by turning the current C# code into Java
- Make NooJ maximally interoperative by making sure it will seamlessly work with major tools

The original idea was to use Java to do this task and this idea was welcome by the IPUP team comprised of experienced Java programmers. However, programmers remained open for other alternatives (MONO framework) since the choice of the implementation framework was strictly a technical issue as long the main objectives are satisfied.

IPUP has made an initial decision to use MONO framework instead of Java after a thorough consideration of pros and cons for each of the alternatives. The brief overview is given in two tables below:

Java framework

Pros	Cons
IPUP team has an extensive experience as Java developers	The members of IPUP team are newcomers in the field of linguistic processing so porting the code to Java can unintentionally create bugs in NooJ engine that would be very difficult to find and correct
Java is an Open Source project	
Java is supported on many platforms	

MONO framework

Pros	Cons
MONO framework supports C#	IPUP team members are newcomers to both C# and MONO
MONO is an Open Source project	
MONO is supported on many platforms	
The author of NooJ prefers to use MONO, because he is familiar with C# and can easily handle the code in the new MONO version of NooJ	

To sum up, both MONO and Java frameworks satisfy the two main requirements defined in DoW related to Open Source implementation and applicability on different platforms. IPUP team inclined to use Java, because the team members have extensive programming experience in Java, while, Max Silberztein, the author of NooJ preferred using MONO. However, what influenced the decision the most was the following: as software developers always tend to minimize risks, so it was assessed that the risk that was to be encountered hard bugs generated by porting NooJ engine to Java is far more severe than the risk of using C# and MONO, which are new to IPUP team members.

The activities performed by the IPUP team related to the implementation of MONO version of NooJ are the following:

- Create the working environment in MONO. IPUP team configured the version control system to support parallel work on NooJ code.
- Create mock-up prototypes. We have created several simple mock-up prototypes to check if NooJ can be successfully ported to MONO framework.

- Create open source file conversion solutions. The .Net version of NooJ used the Microsoft proprietary solutions for converting files in different formats (e.g. html, xml, doc, docx, pdf) to the txt format. However, this was not feasible in the MONO version, therefore IPUP team created the appropriate open source solutions.
- Create new dictionary editor. The dictionary editor in the .Net version of NooJ didn't use a strict grammar for dictionary entries and consequently it couldn't facilitate a precise error reporting. The strict grammar for dictionary entries was developed supervised by Max Silberztein as the author of NooJ and then IPUP team developed a new dictionary parser and the corresponding dictionary editor. The new dictionary editor supports coloring of dictionary entries and precise error reporting. This new dictionary editor was then fully integrated in NooJ.
- Porting NooJ to MONO framework. Microsoft is the owner of many patents related to the implementation of .Net platform, therefore the MONO framework, which supports C#, was actually implemented from scratch and consequently the functionality and behavior of many GUI controls and classes differ from the original .Net controls and classes. Due to these differences, porting NooJ GUI to MONO platform proved to be a difficult task, and required a thorough examination of the complete existing code and performing changes where it was necessary. This activity was the most labor intensive.
- Extensive testing of MONO version of NooJ on multiple platforms. MONO version of NooJ was thoroughly tested by IPUP team on three major platforms: Linux, Mac OSX and Windows. Although the same MONO framework was used, apart from bugs detected on all platforms, some platform specific bugs were also detected.
- Debug MONO version of NooJ on multiple platforms. The majority of bugs detected were related to the MONO implementation and behavior of GUI controls which sometimes differ substantially from the .Net GUI controls. The behavior of MONO GUI controls is not documented anywhere, hence it was extremely difficult to correct these bugs.
- Create installation scripts and installation documentation. It is not enough to install the core MONO framework to be able to start the MONO version of NooJ. IPUP team had to determine which packages are missing and to install them together with NooJ itself for each major platform. Brief installation instructions are also given in the corresponding documentation.
-

All these activities were performed in close collaboration with Max Silberztein as the author of NooJ, engaged by the CESAR project as external expert.

All problems encountered when porting NooJ to MONO framework were technical problems which were solved by investing the adequate effort. However, the real problem emerged with the announcement that Novell will no longer support the MONO project and, as a consequence of this news, that some important platforms (Ubuntu) will no longer support MONO in their new versions. These decisions endangered the ability of MONO framework to comply with the requirement of multiple platform support, which is actually the essential requirement for porting NooJ. Therefore, the events outside the project induced a decision made by Pupin team fully approved by Max Silberztein, as the author of NooJ and Tamás Váradi, as CESAR Project Coordinator, to quit the efforts of porting NooJ to MONO and get back to the initial idea of porting NooJ to Java. This decision was also approved by Hanna Klimek as the responsible project officer at the time.

Cross national collaboration

The effort of this task 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 part of this task 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 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. Based on the agreed guidelines, Partners collected and submitted all relevant metadata to the META-SHARE server and successfully contributed on collecting and publishing all the three upload batches.

The IPR-related tasks were gathered around the following instances, with the main aim of promoting the use of open data and following the Creative Commons and Open Data Commons principles were the main guidelines of the work. Activities carried out in the project involved:

- PSP wide negotiations and discussions on IPR in general and on the proposed license templates
- Clean IPR for all resources involved in upload batches, arrange deposition agreements for resources coming from outside the consortium
- Take an active part in the elaboration and checking of MS-NoRedistribution license template family
- Promote the use of CC or MS licenses, considerably reduce the dominance of CLARIN licenses within CESAR
- Elaborate and implement license solutions and scenarios for often emerging problematic cases
- Work in close and continuous collaboration with other PSPs and META-NET, sharing of experiences, further emerging needs, problems, etc.
- Work in collaborative manner with other LRT projects that are solving their IPR issues in parallel to CESAR activities (e.g. CLARIN, ACCURAT, LetsMT! etc.)
- Apply the most appropriate and the openest possible license scheme out of the set of templates
- Resources resulting from WP3 were made compliant with the legal principles and provisions established and/or completed/amended by the consortium and accepted by the respective right holders.

Within cross-national collaboration a special effort was placed on building a digital exchange platform of META-SHARE nodes. For batch 3, at least one node was created for each language, 7 Partners out of 9 run their own META-SHARE node, and are committed to run them beyond the end of the project. Detailed tasks of the cross-national collaboration and their achievements during the two years of the project are the following:

- Complied with the META-NET recommendations Partners used the META-NET software solutions to implement digital repositories where metadata and/or data are stored or referenced.
- Set up one official CESAR managing node and 6 network nodes for META-SHARE
- During software development phase, run several other nodes in development mode, test and comment on software, help in bugfix and in further development (especially in implementing metadata export/import facility)
- Contribute the resources resulting from WP3 to the META-SHARE pool. Their physical location and 'hosting' was resolved
- Backup all resources by managing and/or network nodes
- Populate metadata descriptions around all official central META-SHARE nodes (as currently they are unsynchronized).

Outreach, awareness and sustainability

In the project, not only technology related work was carried out, but an intensive effort was also placed on disseminating, promoting the outcomes and on preparation of a long-scale sustainability beyond the end of the EU-funded phase. 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.

Dissemination of information about the project was and still is one of activities that rises the awareness about the project itself, its goals, achievements, partners involved as well as funding part(ies). Although it does not produce tangible results that are produced by the main research activities of the project, the Outreach, Awareness and Sustainability work package (WP5) with its results that convey information about the project is considered as important as other WPs.

The overall goal of this task is to disseminate project results and to transfer the project knowledge, technologies, lessons learned and best practices to interested communities and thus to ensure their national, European and global impact and sustainability beyond the project duration.

At the preparation phase of the project we have defined that outreach and awareness activities for CESAR project should be focused on the following target groups:

- Scientific and research community – researchers in the areas of corpus linguistics, computational linguistics, natural language processing, speech processing and language technologies in general;
- Language industry and other business sectors – primarily translation and localization industry companies and professionals, information brokers (documentarians, archivists and digital librarians), (multimedia) language content and service providers (publishers, broadcast companies, news agencies and portals), social media etc. as potential users of language technologies interested in improving the quality of their products when it comes to the consortium languages;
- Society, government and other public decision makers – local governmental officials and industry leaders in this part of Europe.

Beside the general communication channels, for each of these target groups different types of outreaching activities were planned.

Since different target groups would react differently to our messages depending on the content that is ready for presenting, we established three periods in project duration with different priorities regarding the target groups addressing.

Period	Target group	Actions/Instruments
M1-M8	Research community, industry	Papers and presentations with introduction to CESAR project, consortium members, planned actions, expected results
M9-M12	Industry, media, communicators, bodies of language communities, professional LT societies (national level)	Disseminating publications, press releases, announcements, video lectures
M13-M24	Public administration officials (EU and national level), industry, research community	Road show events aiming at mobilisation of national policy, industry and research stakeholders

Table 3. Periods and priorities of the project

At M12 we can say that on national as well as international level the target group "Scientific and research community" has been outreached and has become well aware of META-NET on EU-level and CESAR on regional level. This is not only the result of dissemination activities, presentations and attendance on conferences, but also the result of delivering the resources as META-SHARE platform has grown mature. At this moment that LR&T infrastructure is gaining its desired robustness and assumes the role of pivotal player at EU-level for LR&T.

At M12 it is already the time to turn our focus towards LT users (translators, localisators, media monitors, information brokers etc.) and policy makers (public administration officials at EU and national level). Although by original plan this "turning point" had to happen by the beginning of M9, the delay in META-SHARE proper forced us to wait until the functioning and demonstrable prototype was in place.

At the top of the sustainability actions of the CESAR project is the long term sustainability of the META-SHARE nodes, which equals to the sustainability of the main results, the delivered language resources and tools.

To ensure sustainability of the technical resources developed by CESAR, we propose the following organization of support:

- all partners will be responsible for maintenance of resources and tools provided by their organizations and will thus be appointed as META-CENTREs: institutions administering, supporting, updating and ensuring permanence (including backup) of their resources,
- selected partners will maintain META-NODEs, i.e. the META-SHARE applications functioning as CESAR-related points of entry for requests to access descriptions of META-NET resources and tools synchronized with other META-SHARE nodes; each partner establishing their META-NODE will be responsible for maintenance of the server, applying bugfix releases and updates received from META-SHARE, providing backup of the application and data, monitoring service availability and performance etc.
- one dedicated partner will establish a single point of entry for questions related to CESAR resources and tools in the form of an e-mail address similar to META-SHARE helpdesks (e.g. helpdesk-cesar@example.org) and will be responsible for redirecting questions to respective partners.

The META-CENTREs and the META-NODE will be maintained in 24/7 hour mode. The META-NODEs will be provided by IPIPAN, FFZG, HASRIL, IBL, UBG, LSIL and ULODZ.
The CESAR resource helpdesk will be maintained by IPIPAN with the contribution of other partners in the consortium.

The main goal of sustainability is to ensure prevention from: a) a disconnection to the availability of language resources; b) a duplication of work directed to creation of language resources due to the lack of availability, access or information.
The CESAR consortium has concentrated on all features of language resource that can contribute and have an impact on their sustainability (understood as future availability and usage). The consortium set up a number of requirements in order to meet the sustainability of language resources.

1) Language resource are carefully selected – The consortium selected the best possible resources that will be needed by different groups of end-users. The approach for language resources and tools selection was based on a number of indicators (General evaluation of resources, Total Point Value, Language White Papers) where each language resource was specified according to different groups of criteria. The goal was to ensure as accurate measurement as possible for different quality and quantity parameters. The general evaluation of resources was carried out in three directions: resource upgrade, extension, and cross-lingual alignment.
2) Particular actions are performed to ensure quality and quantity of the selected resources – the selected language resources and tools were upgraded, extended and cross-linked in order to facilitate their usability, professionalism.
3) Language resources are made visible and accessible – Sustainable availability of identified language resources is directed to overcome the restrictions over the public accessibility caused by different personal, privacy or property rights reasons as well as the practice to report on language resources in research publications not providing detailed description and evaluation data for them. Providing exhaustive metadata (both technical and descriptive) enables the users to understand the structure, content and main applications of a resource. The CESAR supported the goal of a common and shared resource description between the four projects constituting META-NET. The META-SHARE metadata are descriptions of Language Resources, encompassing both data sets (textual,

multimodal/multimedia and lexical data, grammars, language models etc.) and tools/technologies/services used for their processing.

The potential impact and the main dissemination activities and exploitation of results

Dissemination activities within the project

Dissemination of information about the projects is one of activities that rises the awareness about the project itself, its goals, achievements, partners involved as well as funding part(ies).
Dissemination activities were pursued in several channels on different occasions.
The channels of dissemination covered mainly traditional dissemination means (flyers, posters, public web-page, scientific / professional conferences / events / lectures / presentations / demonstrations, press releases).

While the project lifetime partners used several channels for coordinated dissemination work. The main channels used for dissemination purposes were:

- webpage – internal and external webpages
- posters – posters presented on conferences
- presentations – presentations presented on conferences and meetings
- fact sheet – official fact sheet used for EU-level dissemination
- flyer – coordinated META-NET flyer localisation to CESAR languages
- publications on Cesar – publications composed at EU-level and at national level
- video lectures – presentations and tutorials made from the results of the project

It may be said with confidence that the reception of project's individuality and uniqueness among all other projects was partly assured through the clearly defined and applied visual identity rules. Defined background and colouring, logo of the project, typefaces used in documents and web page etc. have been used rather consistently whenever a CESAR partner had any kind of presentation of the project. Even the combination of visual identities of META-NET proper with CESAR's own visual identity was successful.

Beside the general visual identity elements (usage of colours, typefaces, etc.), the following dissemination materials were produced:

- official logo: used in all dissemination materials
- official template for presentation (PPT, PPTX and Beamer): used by project partners for presentations at conferences and other events
- QR code: for making simple and efficient shortcut to the URL of the CESAR project website
- deliverables template: used for producing deliverables

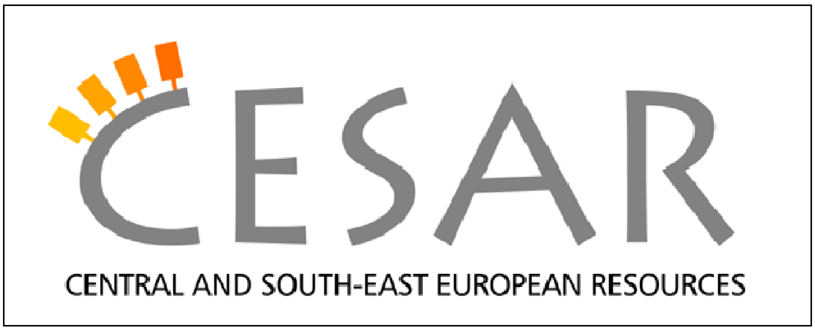


Figure 2. Official CESAR logo

Figure 3. CESAR QR code

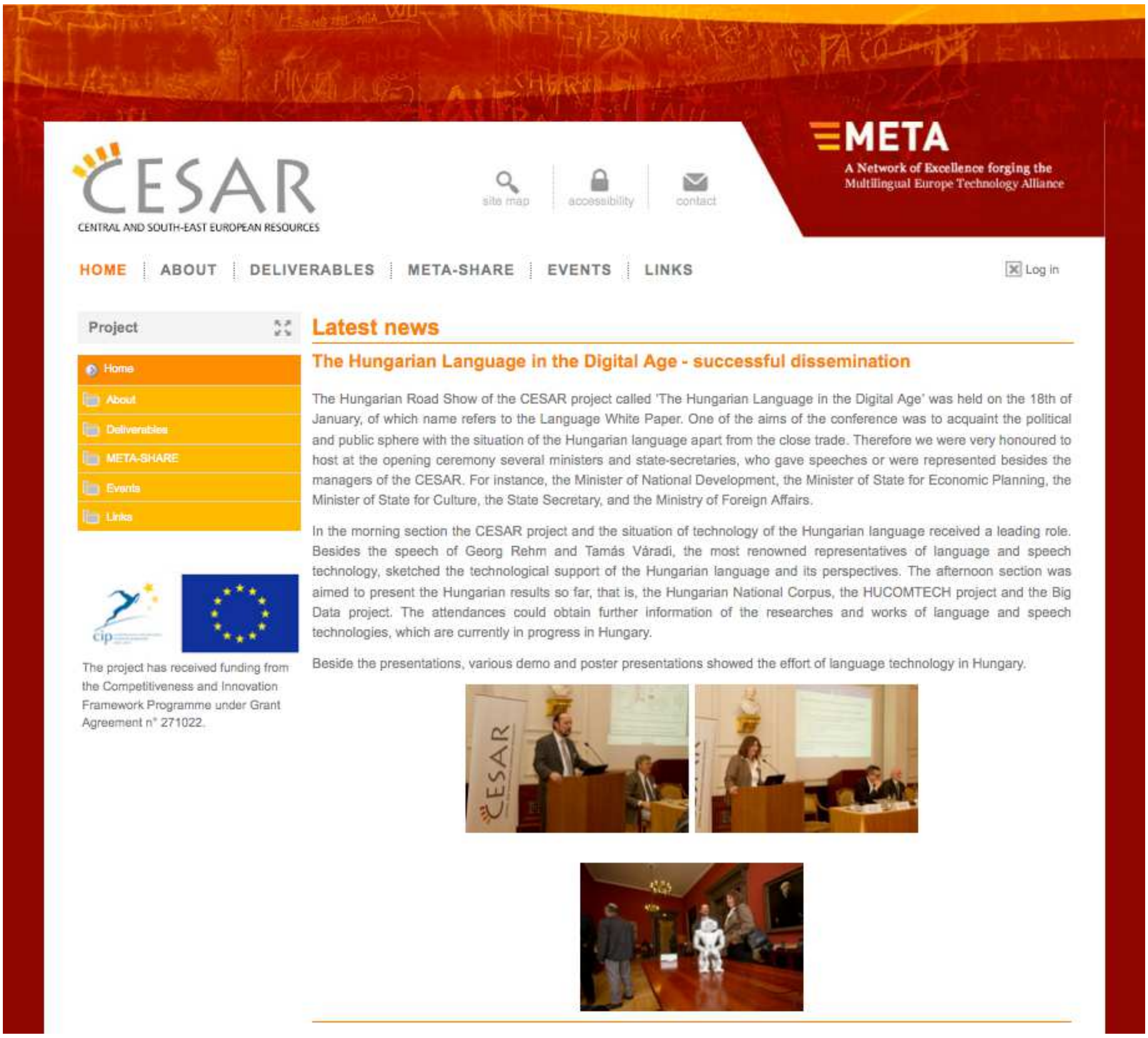
Website

The functional specification for public web site was produced and applied to a web site design. It includes two types of web page elements within the public part of the web site:

- static elements
  - navigation bars (left and top);
  - fixed pages: about the project, list of partners, links, members login, META-SHARE;
- dynamic elements (elements that gained content during the project)
  - current news (directly on the homepage);
  - list of events;
  - list of deliverables;
  - list of documents;

All public results of the project will continue to be published at the website. The graphic design was done by a web designing company (screen-shot of a web design can be seen below) following the recommendations and visual elements from META-NET, but also introducing some specific visual elements according to the defined CESAR visual identity.

The public web site is a dynamic and growing entity and new sections and functionalities will be added through iterative releases and updates. As the project progresses, newer versions of the website will extend these features to CESAR members and the targeted audience.



The project has received funding from the Competitiveness and Innovation Framework Programme under Grant Agreement n° 271022.

Latest news

The Hungarian Language in the Digital Age - successful dissemination

The Hungarian Road Show of the CESAR project called 'The Hungarian Language in the Digital Age' was held on the 18th of January, of which name refers to the Language White Paper. One of the aims of the conference was to acquaint the political and public sphere with the situation of the Hungarian language apart from the close trade. Therefore we were very honoured to host at the opening ceremony several ministers and state-secretaries, who gave speeches or were represented besides the managers of the CESAR. For instance, the Minister of National Development, the Minister of State for Economic Planning, the Minister of State for Culture, the State Secretary, and the Ministry of Foreign Affairs.

In the morning section the CESAR project and the situation of technology of the Hungarian language received a leading role. Besides the speech of Georg Rehm and Tamás Váradi, the most renowned representatives of language and speech technology, sketched the technological support of the Hungarian language and its perspectives. The afternoon section was aimed to present the Hungarian results so far, that is, the Hungarian National Corpus, the HUCOMTECH project and the Big Data project. The attendances could obtain further information of the researches and works of language and speech technologies, which are currently in progress in Hungary.

Beside the presentations, various demo and poster presentations showed the effort of language technology in Hungary.

Paper publications

The initially created flyer and poster were successfully produced and used on different occasions. Several types of CESAR poster and flyers were produced during the project – mostly made in localized versions.

Other dissemination material

Two versions of T-shirts were produced in order to raise the awareness about the project at the conferences and other occasions. Their distribution started at LTC2011 conference where they gained a lot of attention (100 in white and 100 in black).

Producing and distributing the CESAR coffee cup was very successful at LTC2012 conference and this gave us ideas to use this model for widening the spectrum of gadgets available. Pens with pen-drive and note-books , as well as conference bags labelled with CESAR were also produced.

Participation in key conferences

The project continued to disseminate towards the national, EU and the global research community by presenting project results at conferences and workshops and by publishing them in conference proceedings.

Event	Date	Target group	Dissemination activity
GALA2012	March, 2012		

EACL2012	April, 2012	LR&T and NLP researchers, SMEs and large companies in LR&T and NLP field	Oral presentation / poster, demo, contribution to discussion at meetings, flyer / poster distribution
LREC2012	May, 2012	LR&T and NLP researchers, SMEs and large companies in LR&T and NLP field	Oral presentation / poster, demo, contribution to discussion at meetings, flyer / poster distribution / workshop organisation
NooJ2012	June, 2012	LR&T and NLP research community of developers and users of NooJ NLP processing tool	Oral presentation / poster, demo, contribution to discussion at meetings, flyer / poster distribution
TKE2012	June, 2012	LR&T, NLP and semantic web research community	Oral presentation / poster, demo, contribution to discussion at meetings, flyer / poster distribution
ITI2012	June, 2012	ICT and LR&T research community	Oral presentation / poster, demo, contribution to discussion at meetings, flyer / poster distribution
FASSBL 2012	October 2012	LR&T and NLP research community in South Slavic and Balkan countries	Oral presentation / poster, demo, contribution to discussion at meetings, flyer / poster distribution

Table 4. Non-exhaustive list of preferred international key research community conferences that CESAR is expected to appear in 2012 with contributions/presentations/demos.

**Beyond the scientific circles an important target group are players in the language industry, e.g. translation and localization industry. To reach the more key industry players we will address them in coordination with META-NET through the professional organizations such as LISA (Localization Industry Standards Association) and TAUS (Translation Automation User Society), Globalization and Localization Association (GALA) or their respective branches. In fact, contacts with GALA has been established and CESAR was able to exhibit its dissemination materials at eg. at GALA2012 conference.

CESAR organized events – CESAR Road-shows

Beside the activities targeted to research community, the most important means of enhancing awareness in different communities, i.e., business, society and government was a series of nationally organized high-level awareness events („road shows“) that took place in each country in the project duration. We found this form very suitable for local governmental officials and industry leaders in this region of Europe for getting acquainted with the CESAR project, META-NET NoE and the role of LRT in general.

The format of this awareness rising events was a full day gathering where foreign experts (consortium partners from other countries, partners from META-NET, officials from DG Information Society and Media, and experts from other projects from the Call 4 Objective 6.1) gave presentations in order to reach the policy makers and funding agencies at the highest level. Also, leading local industry players were present and demonstrated their products, while researchers presented the current European projects they are involved in. IN the event there were also panel discussions about the future developments of LT for respective language and how META-NET can contribute to that. Also a local Language Whitepaper and the Strategic Research Agenda for Multilingual Europe 2020 was presented. Presenting Language Whitepapers to representatives of government bodies and language policy makers pointing out that prominent national scientists have been involved in drafting the documents and that both national and European language bodies support our cause, will help emphasise our message.

The description and format and the tentative schedule for these events is the following:

9:00-9:30	Registration
9:30-10:00	Opening and introductory speeches by minister of science and/or minster of economy and/or minister of administration, president of the research council and/or economic council/agency, president of academy and/or rector of the university
10:00-10:30	Keynote speaker on LT in general and META-NET in particular with mentioning the Language Whitepaper series + handing over the Language Whitepaper to the highest present government official (suggested speaker: META-NET coordinator Hans Uszkoreit, if possible)
10:30-11:00	coffee break
11:00-11:20	EC InfSo official on the role of LT in multilingual EU and the role of META-NET within (suggested speaker: Robeto Cencioni or Kimmo Rossi, if possible)
11:20-11:40	National/foreign CESAR representative on CESAR and its role in META-NET
11:40-12:00	slot for industry leader
12:00-12:20	slot for government body/agency leader
12:20-12:45	discussion
12:45-14:00	lunch
14:00-14:20	industry/research/government presentation 1
14:20-14:40	industry/research/government presentation 2
14:40-15:00	industry/research/government presentation 3
15:00-15:20	industry/research/government presentation 4
15:20-16:00	coffee break
16:00-17:40	panel discussion (ca 6 participants) on future development of LT for a national language and perspectives for industry on national and EU level (involving national CESAR leader, representatives from the ministries of science, economy, communications, culture, etc., economy chamber/council/agency, etc., leading industry player etc.)
17:40-18:00	general discussion and closing

In parallel a demo session and exhibition of LT products by industrial partners and sponsors, and research projects at both, national and international level was organized.

This regular series of events is considered crucial in the dissemination and outreach actions at each national level. These events were all organized by local organizers, but the logistics was centrally co-ordinated from the WP5 and supported by funds reserved for dissemination to each partner. The target audience was at every event invited, but not limited to, on the basis of the collected internal database of all relevant stakeholders at different national levels.

The CESAR road-shows were organized according to the following schedule:

2012-05-02	Sofia, Bulgaria
2012-06-07/08	Bratislava, Slovakia
2012-09-27/28	Warsaw, Poland
2012-10-29	Belgrade, Serbia
2012-10-30	Zagreb, Croatia
2013-01-08	Budapest, Hungary

This schedule has been derived upon discussion on different aspects that could influence the impact of the event, such as national elections, summer vacations etc. and we see this schedule as the most appropriate. However, it is not strictly fixed and it can be adapted for organizatorial reasons.

On the road-shows conference packages were given for each participant. They include Language Whitepaper on original language or English, mid-term flyer (general or single targeted), t-shirt (S, M, L, XL sizes) and occasionally USB-stick with CESAR logo.

Media appearance

Scientific and other journals

Since our main target group focus was shifted from research community towards industry and policy makers in Y2 of the project, we will not put less effort into publishing papers in the most prestigious scientific journals. Instead, more effort was put to publish articles about CESAR project in professional journals such as *Multilingual Computing Magazine* that addresses professionals in multilingual language industry, or *research*eu* magazine that presents different successful EU-funded research projects. Also, all available relevant national level journals (paper or electronic) were targeted.

Announcements

To draw the attention of the research community to publications and news of the CESAR project, we made more announcements that were published primarily at project public web site. Secondary channel of publishing announcements were different professional mailing lists such as: ACL (acl@aclweb.org), FLaReNet (flarenet_subscribers@ilc.cnr.it), ELRA/ELDA (info@elda.org), ELSNet (elsnet-list@elsnet.org), MT-list (mt-list@eamt.org), CLARIN (members@clarin.eu), CorporaList (corpora@uib.no) and LinguistList for research community.

For industry and policy makers on the national levels there were targeted many potential recipients of CESAR dissemination material such as:

- executive officers of IT companies with R&D activities in the technology domains related to CESAR;
- publishing houses, archives, documentation centres, digital libraries;
- journalists from the (local) scientific/technology press.

Our announcements were regularly published through these channels as well.

Social networks

Social networks presence was coordinated with META-NET proper because it might lead potential target audience to a confusion if two (or even four) projects convey the same short messages. We considered the best practice if the social networks engineering was and is coordinated from the META-NET centre.

Web presence and contact details
Web address of the project:

www.cesar-project.net

Contact persons of partners:
HASRIL: Tamás Váradi (varadi.tamas@nytud.mta.hu)
BME-TMIT: Géza Németh (nemeth@tmit.bme.hu)
FFZG: Marko Tadić (marko.tadic@ffzg.hr)
IPIPAN: Adam Pzrepiorkowski (adamp@ipipan.waw.pl)
ULodz: Piotr Pezik (piotr.pezik@gmail.com)
IBL: Svetla Koeva (svetla@dcl.bas.bg)
UBG: Dusko Vitas (vitas@matf.bg.ac.rs)
LSIL: Radovan Garabík (garabik@kassiopeia.juls.savba.sk)
IPUP: Sanja Vranes (Sanja.Vranes@institutepupin.com)

Németh Géza <nemeth@tmit.bme.hu>, Marko Tadic <marko.tadic@ffzg.hr>, Adam Przepiórkowski <adamp@ipipan.waw.pl>, Piotr Pezik <piotr.pezik@gmail.com>, Svetla Koeva <svetla@dcl.bas.bg>, Dusko Vitas <vitas@matf.bg.ac.rs>, Radovan Garabik <garabik@kassiopeia.juls.savba.sk>

The forground of the project is offered through the channels of META-SHARE infrastructure, thus for each covered countries (project languages) were META-SHARE nodes set up. The nodes offering all covered, CESAR branded language resources and tools can be reached at the following addresses:

- HASRIL node: http://metashare.nytud.hu/
- BME-TMIT node: http://metashare.tmit.bme.hu/
- FFZG node: http://meta-share.ffzg.hr/
- IPIPAN node: http://nlp.ipipan.waw.pl/metashare/
- ULodz node:http://metashare.ia.uni.lodz.pl/
- IBL node: http://metashare.ibl.bas.bg/
- UBG node: http://meta-net.matf.bg.ac.rs:8080/metashare/
- LSIL node: https://metashare.korpus.sk/

Project logo, diagrams, photographs illustrating and promoting the work of the project

Web presence

A focus on the web presence was of particular importance in the dissemination process and our public web site plays the main role in this respect. Beside the general static information and publicly available deliverables, CESAR web site also has several innovative means of dissemination:

- Video lectures: where available, project presentations were digitally video recorded and made viewable with accompanying slides;
- Special areas in CESAR web site with information for different target groups
  - Media: announcements, flyers and posters in PDF;
  - Researchers: upcoming events, project publications;
  - Industry: announcements, demos;
  - General Public: list of Q&A covering the most expected points of interest, a multilingual glossary of language technology;
- NooJ video tutorials: a series of 13 short video clips that contain concentrated explanations of designated terms/points/problems that might appear while using NooJ.

Video lectures

Web-based video lectures represent a chanell of disseminating information about the project that has no time and/or space limitation. They are available 24/7 for anyone having access to a modest computer. In this way the ideas and results covered by CESAR can reach around the globe. The following talks by CESAR participants were video recorded and are available at following URLs:

1. **Bulgarian road-show**, Sofia, Bulgaria, 2012-05-02
   a. the whole road-show was video recorded
2. **LREC2012**, Istanbul, Turkey, 2012-05-21/27:
   a. NooJ LREC Tutorial (half day tutorial)
   b. Varadi: presentation of the CESAR project
3. **Slovak road-show**, Bratislava, Slovakia, 2012-06-07/08
   a. the whole road-show was video recorded
4. **META-FORUM2012**, Bruxelles, Belgium, 2012-06-20/21
   a. Andras Kornai: http://videolectures.net/metaforum2012_kornai_language/
   b. Tamás Váradi: http://videolectures.net/metaforum2012_varadi_cesar/
   c. Panel META-NET and beyond: http://videolectures.net/metaforum2012_meta_share_discussion/
5. **Polish road-show**, Warsaw, Poland, 2012-09-26/27
   a. the whole road-show was video recorded
6. **Serbian road-show**, Belgrade, Serbia, 2012-10-29
   a. the whole road-show was video recorded
7. **Croatian road-show**, Zagreb, Croatia, 2012-11-30
   a. the whole road-show was video recorded
8. **Hungarian road-show**, Budapest, Hungary, 2013-01-18
   a. the whole road-show was video recorded

Promotional materials

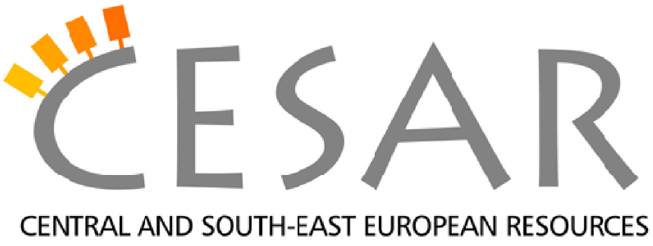


Figure 4. Logo of CESAR-project

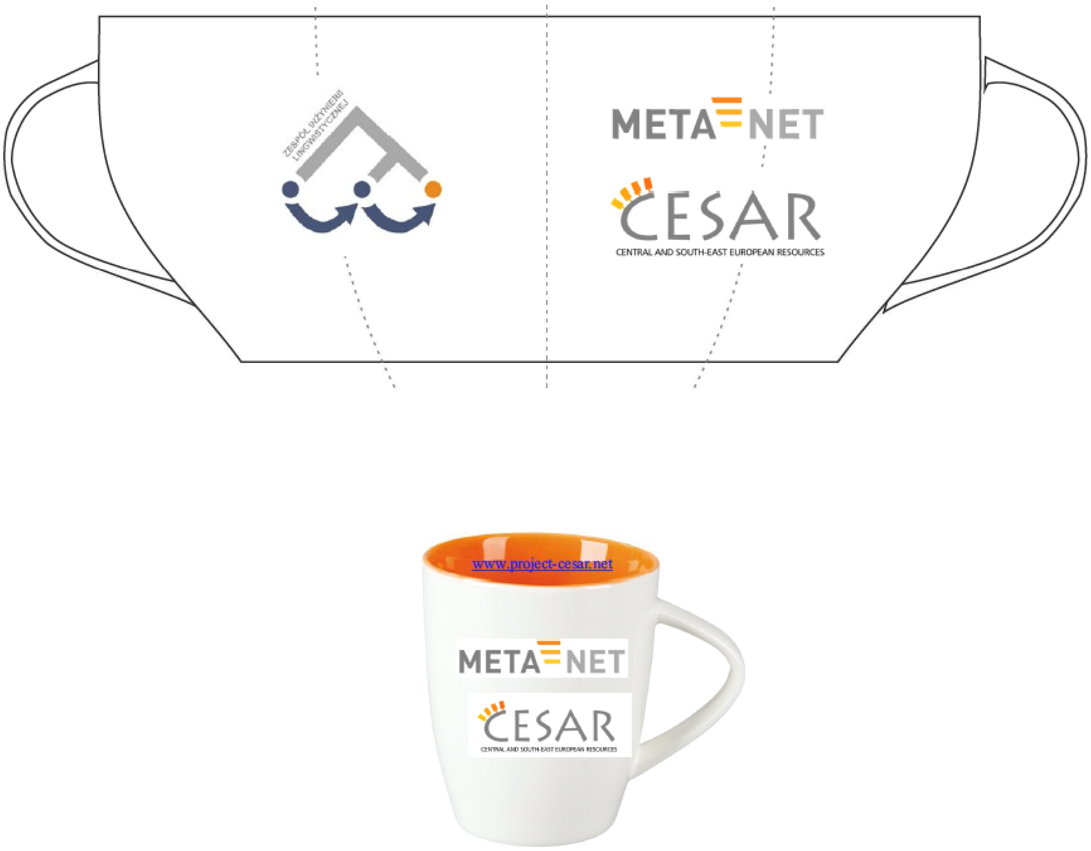
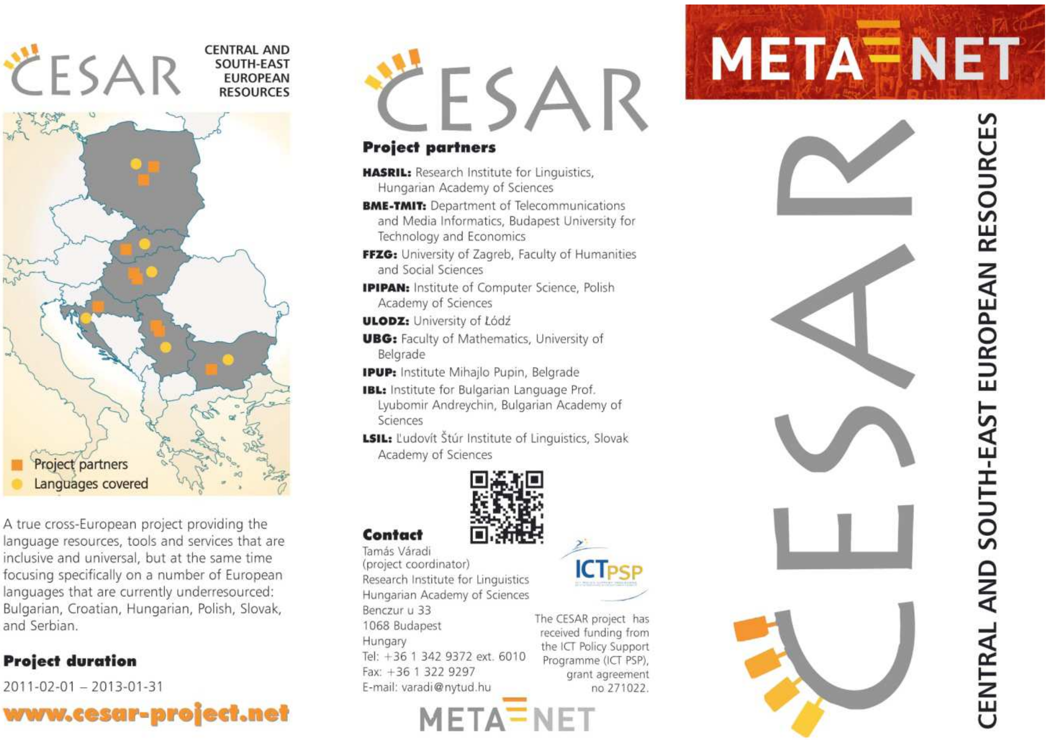

ZESPÓŁ INŻYNIERII LINGWISTYCZNEJ
META-NET
CESAR
CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES
www.project-cesar.net
META-NET
CESAR
CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES

Figure 5. CESAR and META-NET visual identity elements on a cup

Figure 6. CESAR tall coffee cup



CESAR CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES

Project partners
Languages covered

A true cross-European project providing the language resources, tools and services that are inclusive and universal, but at the same time focusing specifically on a number of European languages that are currently underresourced: Bulgarian, Croatian, Hungarian, Polish, Slovak, and Serbian.

Project duration

2011-02-01 – 2013-01-31

www.cesar-project.net

CESAR

Project partners

HASRIL: Research Institute for Linguistics, Hungarian Academy of Sciences

BME-TMIT: Department of Telecommunications and Media Informatics, Budapest University for Technology and Economics

FFZG: University of Zagreb, Faculty of Humanities and Social Sciences

IPIPAN: Institute of Computer Science, Polish Academy of Sciences

ULODZ: University of Łódź

UBG: Faculty of Mathematics, University of Belgrade

IPUP: Institute Mihajlo Pupin, Belgrade

IBL: Institute for Bulgarian Language Prof. Lyubomir Andreychin, Bulgarian Academy of Sciences

LSIL: Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences

Contact

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

ICTPSP

The CESAR project has received funding from the ICT Policy Support Programme (ICT PSP), grant agreement no 271022.

META-NET

META-NET

CESAR

CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES

Figure 7. CESAR initial flyer in English

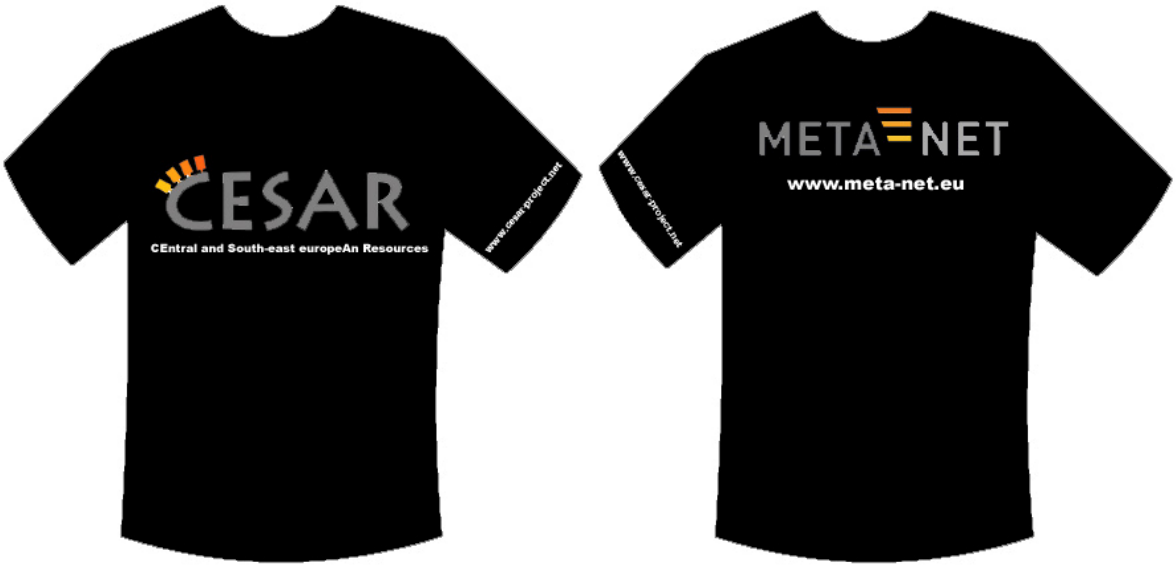


Figure 8. Front and back of a black CESAR T-shirt

Figure 9. CESAR table calendar

Photographs of the events

Foreword
A digital *lingua franca*?
META-NET
CESAR
CESAR
CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES

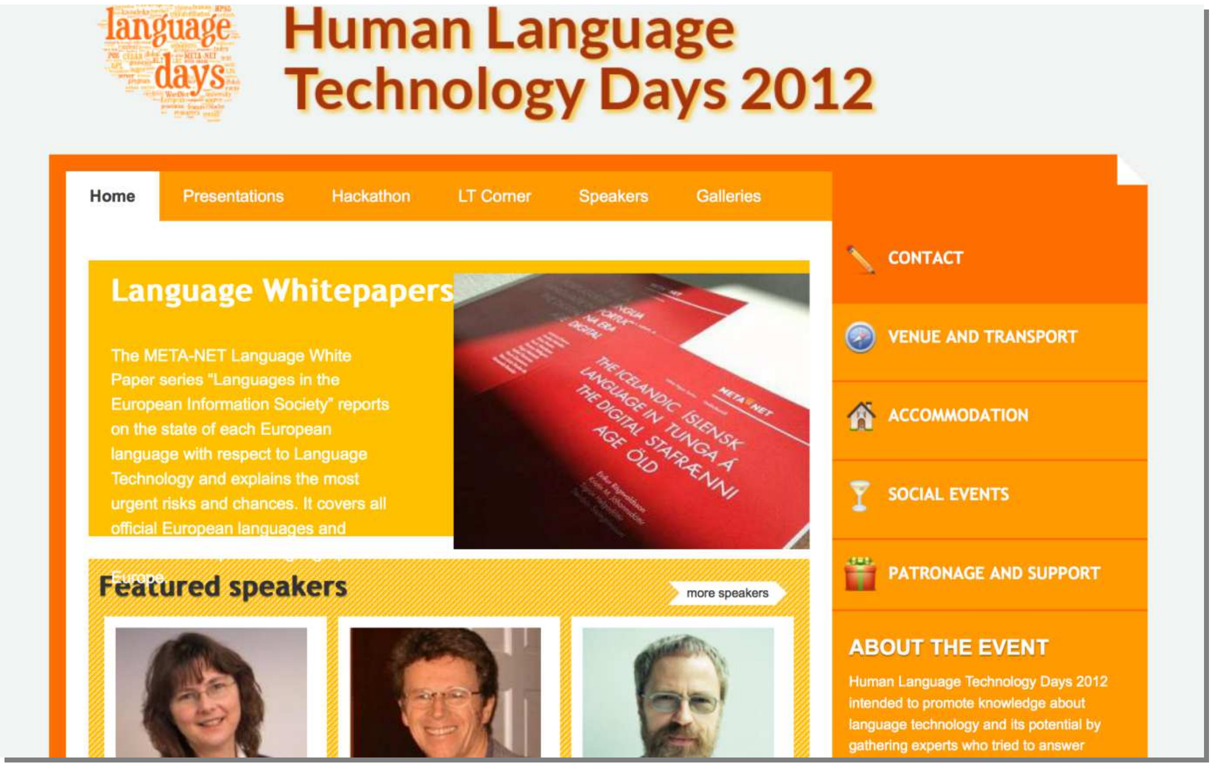
Figure 10. Bulgarian road-show, Sofia, 02-05-2012

Project objectives
META-NET
CESAR
provide a description of the national landscape in terms of
language use, language-savvy products and services, language technologies and resources
contribute to a pan-European digital language resources exchange (META-SHARE)
enhance, extend, document, standardize, cross-link, cross-align resources and tools
mobilise national and regional stakeholders, public bodies and funding
reinvigorate cooperation between key technology partners in the region
collaborate with other partner projects
bridge the technological gap between this region and the other parts of Europe by

Figure 11. Slovak road-show, Bratislava, 07/08-06-2012

Figure 12. Croatian road-show, Zagreb, 30-11-2012

Figure 13. Hungarian road-show, Budapest, 18-01-2013



Human Language Technology Days 2012

Language Whitepapers

The META-NET Language White Paper series "Languages in the European Information Society" reports on the state of each European language with respect to Language Technology and explains the most urgent risks and chances. It covers all official European languages and

Europe

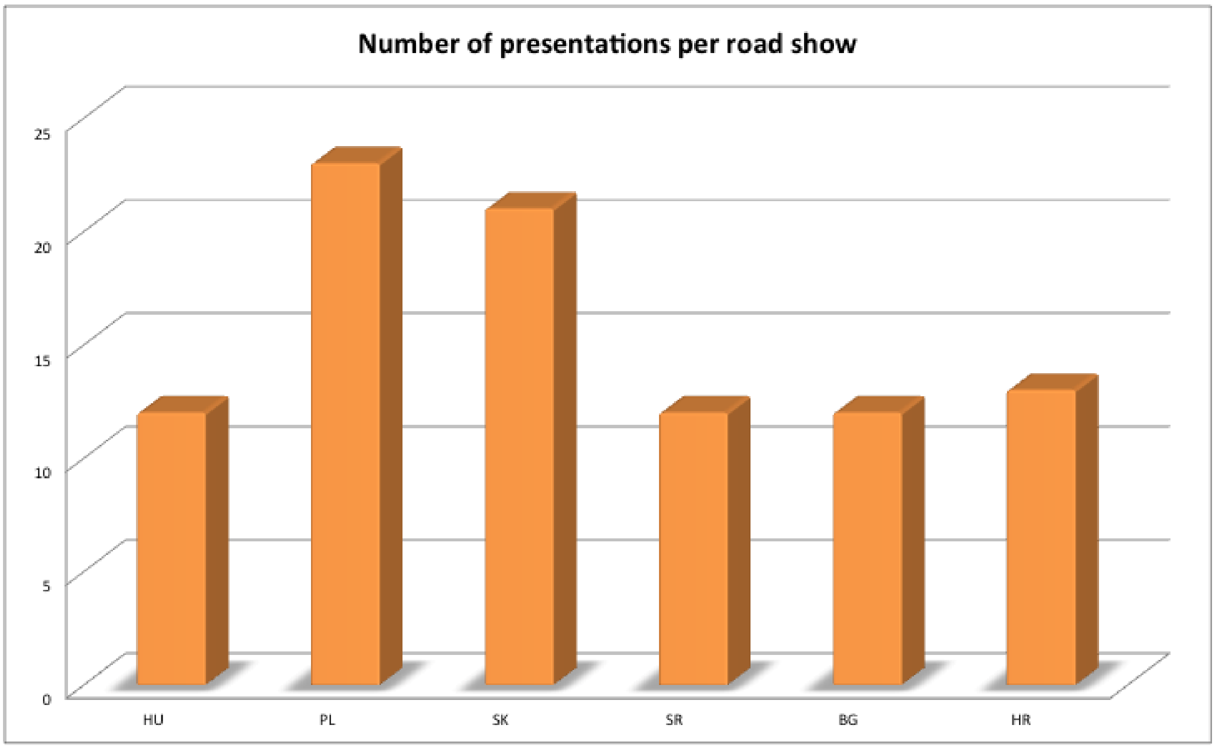
Featured speakers

ABOUT THE EVENT

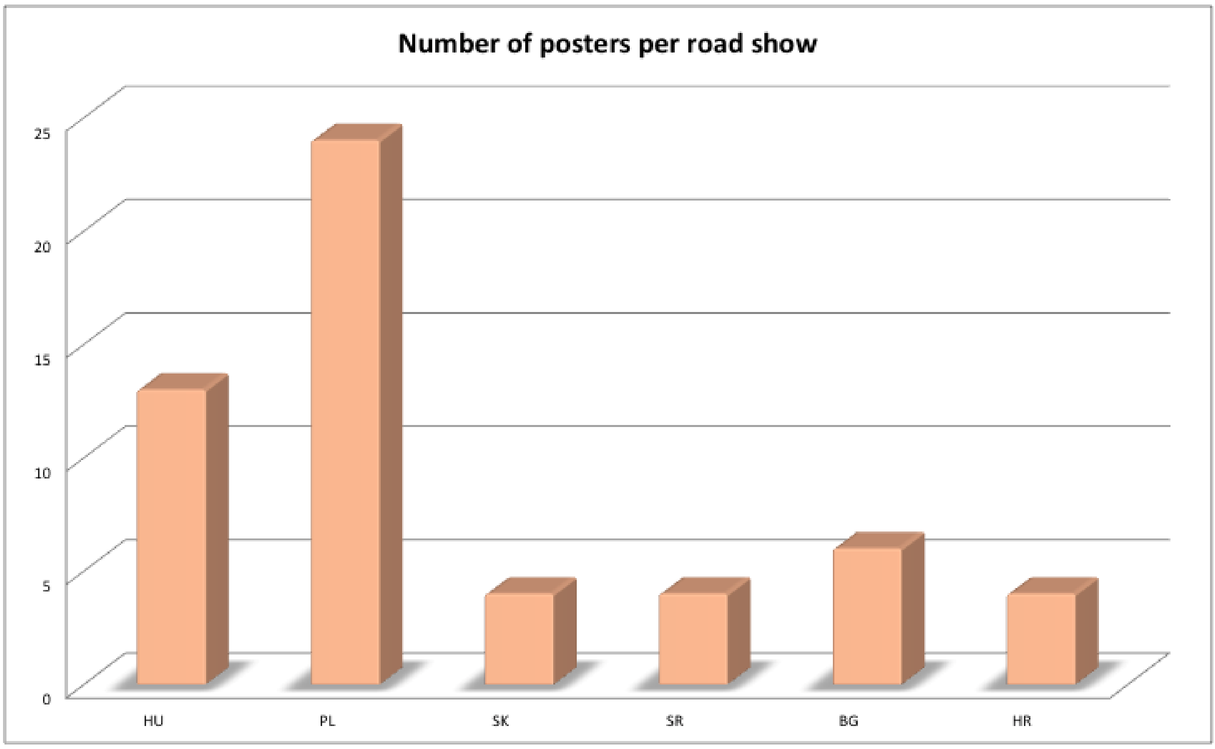
Human Language Technology Days 2012 intended to promote knowledge about language technology and its potential by gathering experts who tried to answer

Figure 14. Web pages of the Polish road-show under the title HLT Days

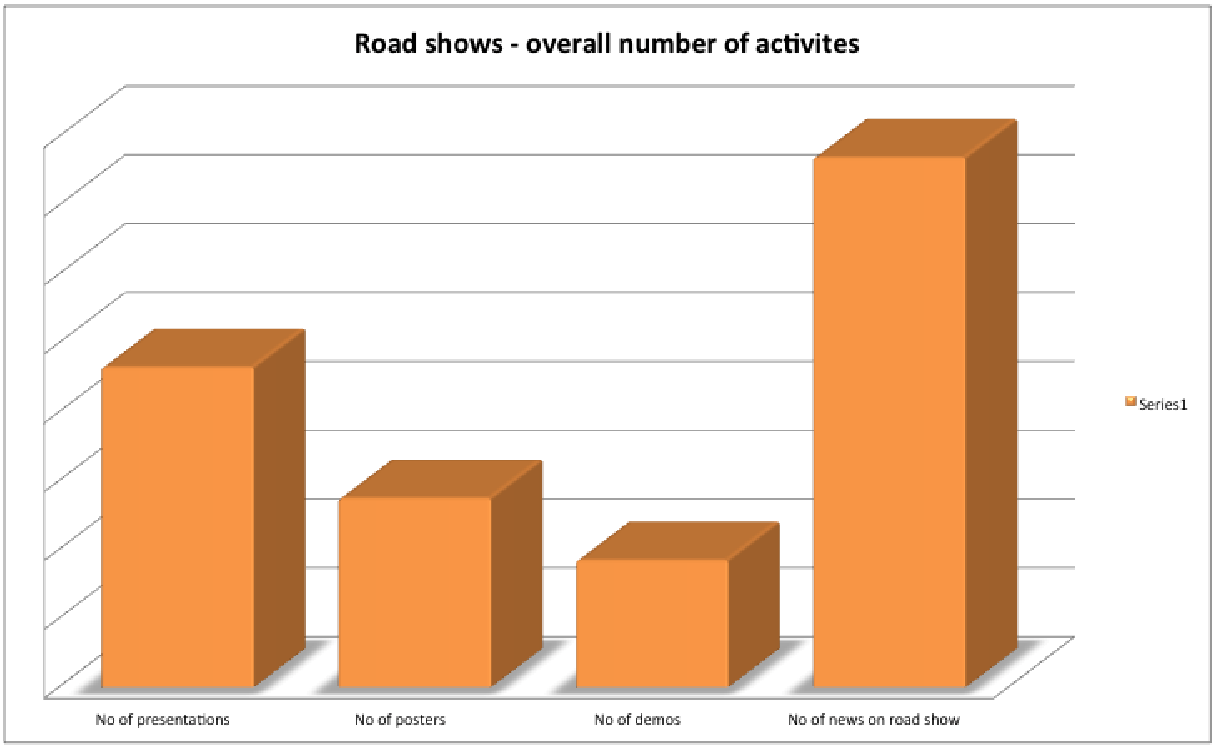
Diagrams of the promotion of the work



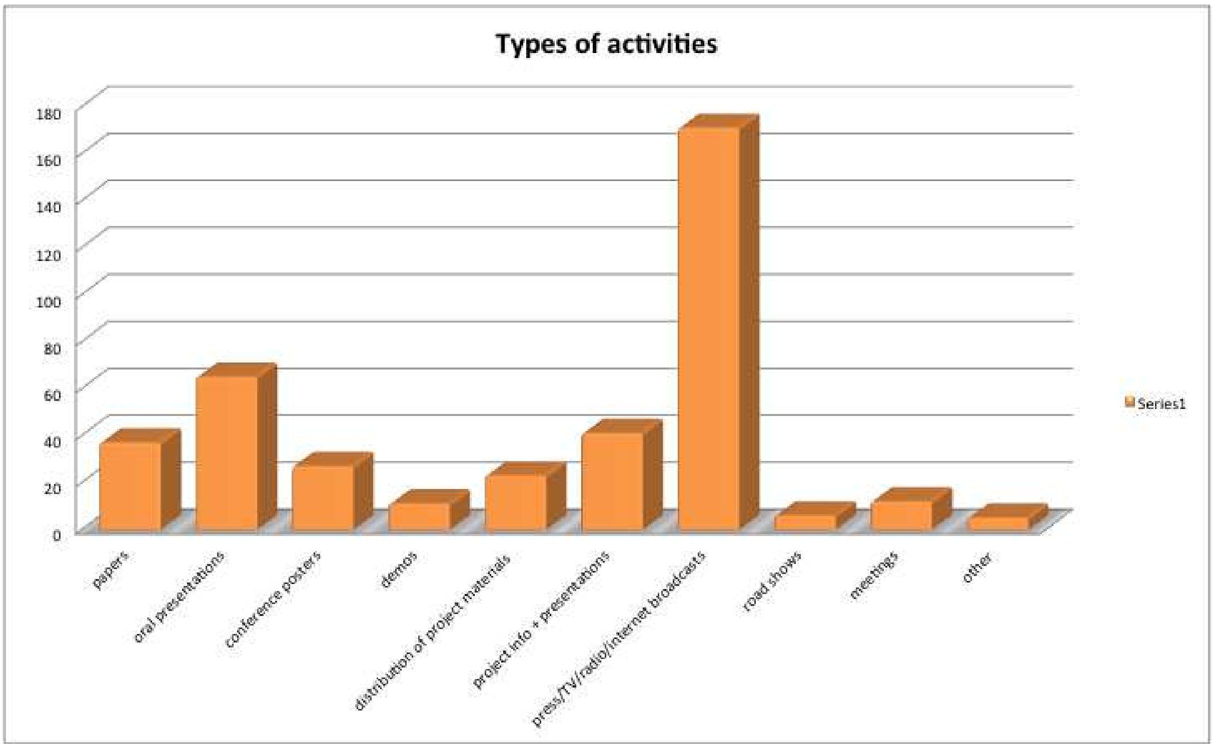
Graph 1. Number of presentations per road-show



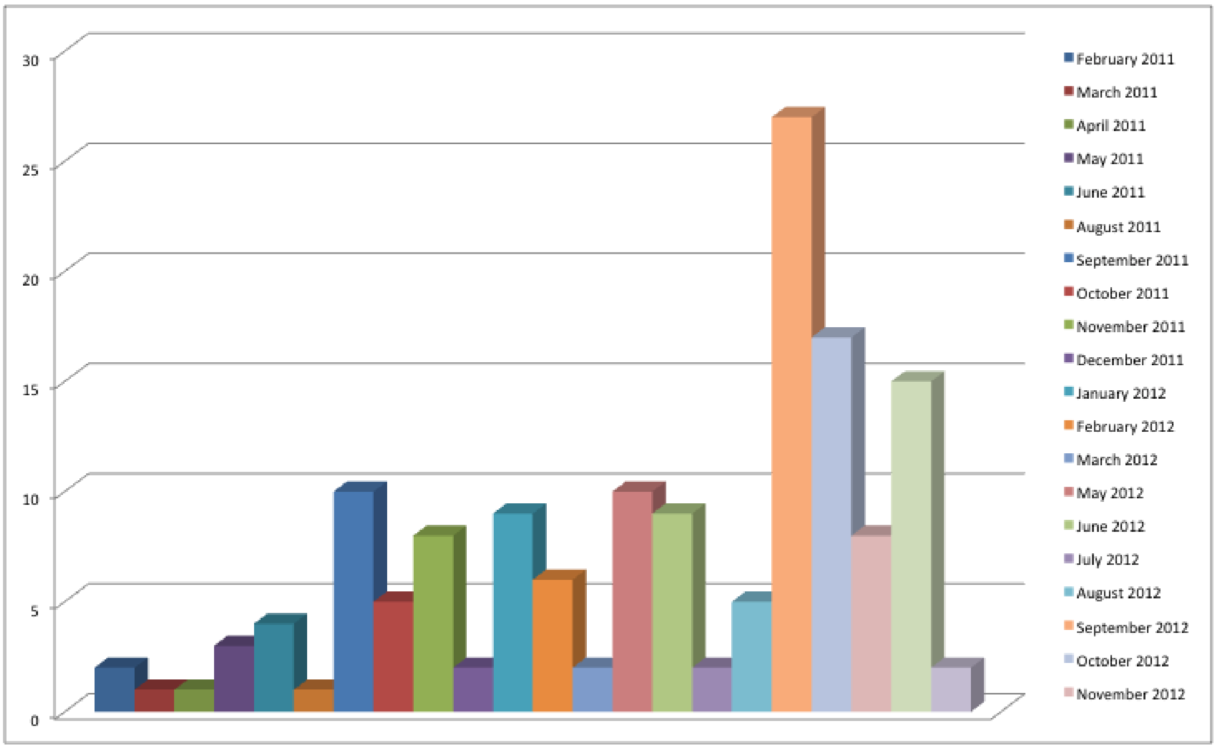
Graph 2. Number of posters per road-show



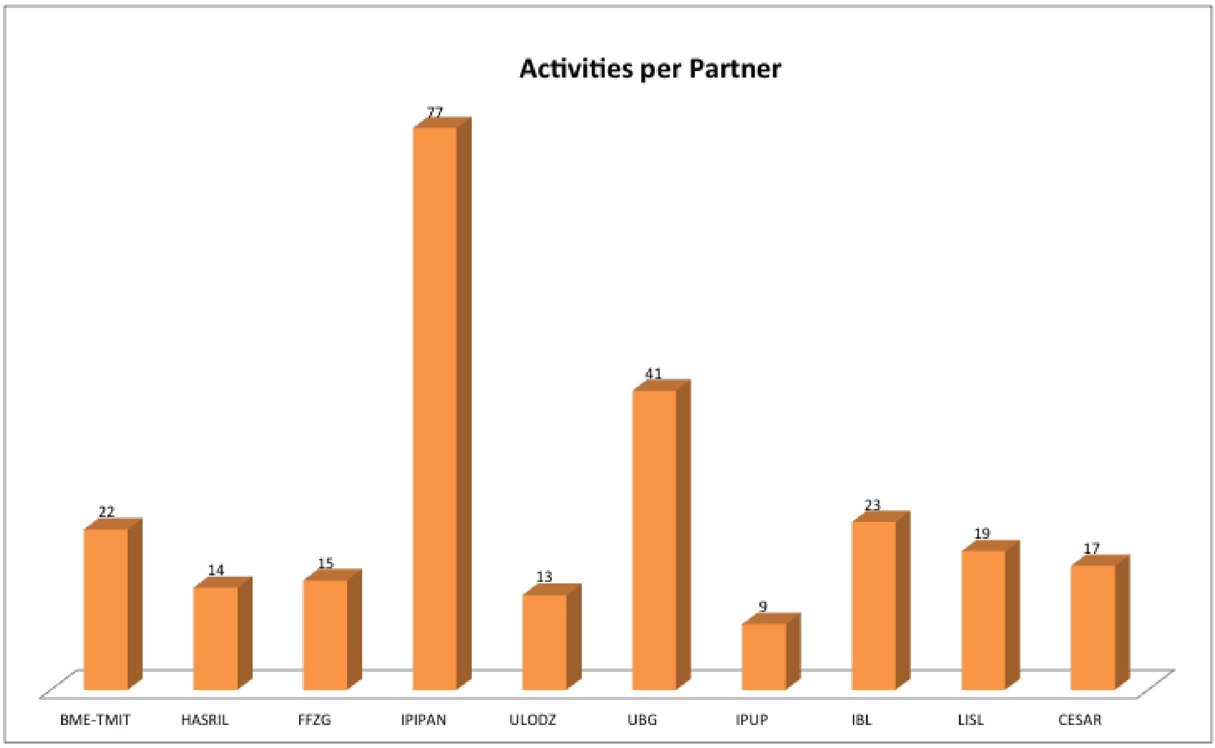
Graph 3. Overall number of activities per road-show



Graph 4. Statistics on types of dissemination activities



Graph 5. Distribution of dissemination activities over months



Graph 6. Distribution of dissemination activities over partners

Use and dissemination of foreground

Language resources and tools enriched within the CESAR project will be promoted and hosted through the META-SHARE infrastructure and platform. META-SHARE infrastructure relies on the operation of interlinked META-SHARE nodes that are distributed and autonomously maintained by the participating institutions. With the descriptions below CESAR guarantees the maintenance and the sustainability of META-SHARE infrastructures for the proposed years and for countries involved in CESAR project. HASRIL as coordinator of CESAR will ensure sharing and functioning of the CESAR resources through the META-SHARE infrastructure.

To ensure sustainability of the use and usability of the technical resources developed by CESAR, we propose the following organization of support:

- all partners will be responsible for maintenance of resources and tools provided by their organizations and will thus be appointed as META-CENTREs: institutions administering, supporting, updating and ensuring permanence (including backup) of their resources,
- selected partners will maintain META-NODEs, i.e. the META-SHARE applications functioning as CESAR-related points of entry for requests to access descriptions of META-NET resources and tools synchronized with other META-SHARE nodes; each partner establishing their META-NODE will be responsible for maintenance of the server, applying bugfix releases and updates received from META-SHARE, providing backup of the application and data, monitoring service availability and performance etc.
- one dedicated partner will establish a single point of entry for questions related to CESAR resources and tools in the form of an e-mail address similar to META-SHARE helpdesks (e.g. helpdesk-cesar@example.org) and will be responsible for redirecting questions to respective partners.

The META-CENTRE and the META-NODEs will be maintained in 24/7 hour mode. The META-NODEs will be provided by IPIPAN, FFZG, HASRIL, IBL, UBG, LSIL and ULODZ, while the central node is and will be maintained by IPIPAN.

The CESAR resource helpdesk will be maintained by IPIPAN with the contribution of other parnters in the consortium.

META-SHARE functionalities and services to offer and sustain

To keep the common identity of the META-SHARE platform, end-user and system management functionality of the infrastructure provided by the META-NODEs will be entirely based on functionalities offered by the META-SHARE software platform. For external users the access to all CESAR covered language resources will be read-only and will consist of all features that are provided by this software:

- keyword-based LRT search and browse (with standard functionality such as faceted filtering, ordering, paging etc.),
- access to usage statistics (most viewed, top downloaded, most recently updated resources, top queries, latest queries and list of similar LRTs),
- user registration and login (necessary e.g. to download resources),
- downloading resources (if offered by depositor)

To ensure adequate level of service and support after the end of the CESAR project, equipment dedicated to the maintenance of CESAR resource metadata descriptions and backup version of resources was specified and provided. In some cases, e.g. FFZG, the physical configuration was replaced by virtual server, provided by a large high-power computing centre, that can be scaled as needed regarding the number of processors, size of RAM or volume of virtual hard disks and that will take care about all the maintenance (backup) and uptime issues (24/7/365)

Maintenance and support of CESAR supported META-SHARE nodes

Administration and maintenance of the META-SHARE node in the distributed network of core nodes consists of:

- keeping the server up to date with respect to security-related software updates,
- installing bugfix releases of META-SHARE,
- monitoring server availability and performance,
- periodically checking log files (e.g., to verify that synchronization between nodes is still operational),
- fixing any problems on maintenance level (e.g., server crashes),
- maintaining help-desk for the central node as well as for all nodes.

Estimated effort of the administration task by an experienced system administrator is approx. 2 days per month (1 PM per year per META-NODE).

The administration will not cover any META-SHARE application-related implementation, documentation neither bugfixing, which will be provided by META-SHARE.

In the case of configuration of META-NODE on the virtual server(s), some of the services (uptime assurance, backups, server crashes, performance issues, security-related issues, etc.) are taken care by the large computing centre providing virtual server facility, so efforts of administration tasks could even be lower.

CESAR LRT-related support of users will consist of:

- monitoring LRT availability,
- redirecting LRT-related questions to respective partners (resource distributors),
- maintaining communication with partners to track LRT-related issues and build know-how necessary to provide first level of support to external users.

Estimated effort of the support task by an experienced consultant is approx. 0,25 hour per month per resource (with current 251 CESAR resources it makes 0,36 PM per month and 4,27 PM per year).

Commitments of CESAR partners

All countries in CESAR have expressed their willingness to maintain a META-CENTRE or META-NODE with the aim to serve as centre of repositories of LRT-s (Bulgaria, Croatia, Hungary, Serbia, Slovakia and Poland). The commitments of partners to participate in the long time maintenance of the META-SHARE is underpinned with duly signed Letters of Intent in 2012, which can be found in Attachment I of D5.3b Sustainability strategy and plans beyond the end of the project), but also in the fact that all partners willing to set up the META-NODE, have already done so by the end of the project.

Plans for using the CESAR resources

The outcome of the project was (as it is written in the Annex I.) to select the best resources in all covered languages, to make all resources in compliance with the up-to-date requirements of the field and offer resources for all users (possibly for researchers and business usage).

The main goal of sustainability is to ensure a prevention from: a) a disconnection to the availability of language resources; b) a duplication of work directed to creation of language resources due to the lack of availability, access or information.

The CESAR consortium has concentrated on all features of language resource that can contribute and have an impact on their sustainability (understood as future availability and usage). The consortium set up a number of requirements in order to meet the sustainability of language resources.

1) Language resource are carefully selected - a methodology and criteria that allow partners to assess the quality and importance of language resources are established and carefully followed. The aim is to

ensure a balanced coverage of resources for different end users and tasks, groups of products and services.
2) Particular actions are performed to ensure quality and quantity of the selected resources - upgrading, extending and linking the resources, aligning resources across languages.
3) Language resources are made visible and accessible – META-SHARE metadata descriptions are based on established standards, best practice and users needs. Providing exhaustive metadata descriptions enables the users to find out the most suitable resource and to use it in an appropriate way

Language resources were made visible and accessible

Sustainable availability of identified language resources is directed to overcome the restrictions over the public accessibility caused by different personal, privacy or property rights reasons as well as the practice to report on language resources in research publications not providing detailed description and evaluation data for them.
Providing exhaustive metadata (both technical and descriptive) enables the users to understand the structure, content and main applications of the chosen resource. The CESAR consortium supported the goal of a common and shared resource description between the four projects constituting META-NET (i.e., CESAR, METANET4U and META-NORD, and T4ME).

Pomoting the NooJ open source language tool

NooJ, as one of the most popular linguistic tools (not just in LT community), was one of foci of the CESAR project, and, therefore, one of its tasks was devoted to its translation to open source and its porting to multiple platforms – in norder to ensure a wider usability amongst users. The task required cross-national collaboration, with Max Silberztein (France) as NooJ's author on one side and the team of Institute Mihajlo Pupin (Serbia) that was porting NooJ to Java platform on the other. After the end of the project NooJ became become an open source software and many developers from different countries are involved in applying the existing NooJ solutions or in providing various functional extensions to the existing code. The Institute Mihajlo Pupin supports the open source NooJ community by providing maintenance of the existing code and helping open source NooJ developers.
At the time of compiling this report NooJ community made major steps to became an a legal entity (named as NooJ Association). NooJ Association will be an AISBL (Association internationale sans but lucratif - a special non-profit international association under Belgian law).
Another way to ensure wider visibility, accessibility and sustainability of NooJ was a series of NooJ video tutorials explaining basic functionalities and use of NooJ. The resulting video clips were produced and published on YouTube and CESAR web site, replicated to doubly ensure their easy accessibility, thus contributing to the sustainability of the system.

National cooperation in the interest of long term usage

All partners are basic centres of language technology field in their respective country with wide range of contacts in their national scene. The research community at national level was therefore reached by a range of traditional or not so traditional dissemination channels (flyers, posters, public web-page, conferences, events presence, lectures, presentations, demonstrations, press releases, but also video lectures, video presentations and publicity in national media) in order to attract the players to participate in sharing resources and tools through META-CENTRES.

National cooperation strengthening the wider usage of the foreground is also organized through national infrastructure collaborations (which can be domestic of as part of a wider, international presence) such as the CLARIN or its local forms (such as Hun-CLARIN). In Hungary the national cooperation is organized through common platforms such as the Language and Speech Technology Platform which gathers the main actors in LT field ranged from the Academic part, through universities and business partners. This platform was created to map the Hungarian LT community and enhance activity in the proper field. Research community in Hungary is also gained across the Hun-CLARIN platform which is also a good opportunity for dissemination purposes.

In Croatia, the Croatian Language Technologies Society (CLTS) plays the role of hub where LT activities are tracked and, if possible, co-ordinated at the national level. Members of this professional non-profit society are members of all LT-relevant institutions in Croatia and are participants in the national LT projects, funding of which has just been closed on 2012-10-31. Since no new calls are issued at the moment, it is hard to tell how the LT in general, and CESAR-related activities in particular, will be supported nationally in the future since there are no infrastructure projects running at the moment. Members from EU-funded projects will exercise their dissemination activities throughout the projects' duration and some of these activities will have long-lasting effect relevant for CESAR language resources. since they will be used in these projects (e.g. FP7 project XLike).

The CLTS is an co-organiser of a regional bi-annual conference Formal Approaches to South-Slavic and Balkan Languages (FASSBL, http://www.fassbl.org) and this conference serves as the meeting point of researchers from computational linguistics or LT and industrial partners for respective languages where all recent research and project activities were presented. The conference is regularly supported by the Ministry of Science, Education and Sports of the Republic of Croatia and from other sources.

The CLTS also maintains the Croatian Language Technologies Portal (http://jthj.ffzg.hr) that delivers information about LT activities nationally and internationally, and is being online since 2000.

The Human Technology Group, supported by the Faculty of Mathematics and the Faculty of Philology at the University of Belgrade coordinates most of the activities related to language technologies in Serbia. These faculties have a lot of experience in organizing international conferences, a number of them related to LT.

Dissemination and long time dissemination efforts

Dissemination is a process made by all partners in tight cooperation at national and international level. Dissemination activities at the European and global level were coordinated and harmonised with META-NET dissemination activities in order to maximise the impact by controlled spread of different participants at different events. The joint effort of the project participants could have been seen in the joint presence at conferences and workshops of highest level in LT world.

CESAR's perspective is to be visible at main LT events both at national and international level. This basic idea will remain after project ended. The dissemination (showing efforts for sustainability) will be alive mainly in conferences and events, and spread via traditional ways (posters, flyers).

CESAR put an emphasis on the video tutorials as a channel of dissemination will be available long time after the project end. At this moment a series of short NooJ video tutorials describing how to use NooJ was prepared. These video tutorials will remain accessible on-line for the interested audience and they demonstrate a clear connection with the CESAR project through visual elements (logos, web site address etc.). They should be considered an integral part of the NooJ as an open source bundle and will be accessible both at NooJ and CESAR web site.

Also, video lectures of CESAR partners' presentations at different conferences and workshops will remain accessible on-line after the project end to continue CESAR presence.

LIST OF SCIENTIFIC (PEER REVIEWED) PUBLICATIONS, STARTING WITH THE MOST IMPORTANT ONES

NO.	Title	Main author	Title of the periodical or the series	Number, date or frequency	Publisher	Place of publication	Year of publication	Relevant pages	Permanent identifiers[1] (if available)	Is/Will open access[2] provided to this publication?
1	*CESAR resources in META-SHARE repository*	*Garabík, R., Koeva, S., Krstev, C., Ogrodniczuk, M., Pęzik, P., Przepiórkowski, A., Stanojević, M., Tadić, M., Váradi, T., Vicsi, K., Vitas, D., Vraneš, S.*	*Proceedings of the 5th Language & Technology Conference, LTC2011*			*Poznan*	*2011*	*583*		*yes*
2	*Detecting Gaps in Language Resources and Tools in the Project CESAR*	*Garabík, R., Koeva, S., Ogrodniczuk, M., Tadić, M., Váradi, T., Vitas, D.*	*Proceedings of the 5th Language & Technology Conference, LTC2011*			*Poznan*	*2011*	*37-41*		*yes*
3	*Introducing the CESAR Project*	*Váradi, T.*	*INFOtheca - Journal of Information and Library Science*	*12/2*			*2011*	*71-74*		
4	*Towards an open repository of Polish language*	*Pęzik P., Ogrodniczuk M.,*	*Proceedings of 5th Language & Technology*		*Fundacja Uniwersytetu im. A.*	*Poznan*	*2011*	*511-515*		

	resources.	*Przepiórkowski A.*	*Conference*		*Mickiewicza*					
5	*Syntactic Patterns of Verb Definitions in Croatian WordNet*	*Bekavac, B., Šojat, K.*	*Automatic Processing of Various Levels of Linguistic Phenomena: Selected Papers from the NooJ 2011 International Conference*		*Cambridge Scholars Publishing*	*Newcastle upon Tyne*	*2012*	*109-117*		
6	*Applications of Bulgarian-English Parallel Corpus for Exploring Translational Asymmetries*	*Koeva, S., Stoyanova, I. and Dekova, R*	*Automatic Processing of Various Levels of Linguistic Phenomena: Selected Papers from the NooJ 2011 International Conference*		*Cambridge Scholars Publishing*	*Newcastle upon Tyne*	*2012*	*227-240*		
7	*Central and South-European language resources in META-SHARE*	*Maciej Ogrodniczuk, Radovan Garabík, Svetla Koeva, Cvetana Krstev, Piotr Pęzik, Tibor Pintér, Adam Przepiórkowski , György*	*Infotheca*	*12/1*			*2012*			

		Szaszák, Marko Tadić, Tamás Váradi, Duško Vitas								
8	*At Least 21 European Languages in Danger of Digital Extinction; Студија водећих европских експерата за језичке технологије упозорава да већина европских језика неће опстати у дигиталном добу*	*Georg Rehm, Hans Uszkoreit*	*Infotheca*		*Serbian Academic Library Association*	*Belgrade*	*2012*			

9	*Rozvoj jazykových technológií na Slovensku a vo svete (10 rokov Slovenského národného korpusu) / Development of the Human Language Technologies and Resources in Slovakia and in the world (10 years of the Slovak National Corpus)*	*Gajdošová, K.*	*Kultúra slova*	*3*			*2012*	*167-171*		
10	*Српски језик у дигиталном добу -- The Serbian Language in the Digital Age*	*Vitas, D., Popović, Lj., Krstev, C., Obradović, I., Pavlović-Lažetić, G., Stanojević, M*	*META-NET White Paper Series*		*Springer*	*Berlin*	*2012*			
11	*Българският език в дигиталната епоха. The Bulgarian Language in the Digital Age*	*Blagoeva, D., Koeva, S., Murdarov, V*	*META-NET White Paper Series*		*Springer*	*Berlin*	*2012*			
12	*Open source multi-platform NooJ for NLP*	*Silberztein, M.; Váradi, T.; Tadić, M*	*Proceedings of COLING2012*		*ACL*		*2012*	*401-408*		

13	*Central and South-East European Resources in META-SHARE*	*Tadić, M.; Váradi, T*	*Proceedings of COLING2012*		*ACL*		*2012*	*431-438*		
14	*Annotating the Corpus of Contemporary Serbian*	*Utvić, M.*	*INFOtheca - Journal of Information and Library Science*		*Serbian Academic Library Association*	*Belgrade*	*2011*	*36-47*		
15	*Orwell's 1984 – the Case of Serbian Revisited*	*Vitas, D., Krstev, C.*	*Proceedings of 5th Language & Technology Conference*		*Fundacja Uniwersytetu im. A. Mickiewicza, Poznań*	*Poznan*	*2011*	*570-574*		
16	*A formánsmenetek rendszere CVC kapcsolatok magánhangzóiban a C képzési helyének függvényében (Formant trajectory types in vowels of CVC sequences as the function of the articulatory place of Hungarian consonants)*	*Abari Kálmán – Olaszy Gábor*	*Beszédkutatás*	*20*	*MTA Nyelvtudományi Intézet*	*Budapest*	*2012*	*94-106*		

17	*Disambiguation of Homographic Adjective and Adverb Forms in Croatian Texts.*	*Merkler, D., Berović, D., Agić, Ž*	*Automatic Processing of Various Levels of Linguistic Phenomena: Selected Papers from the NooJ 2011 International Conference*		*Cambridge Scholars Publishing*	*Newcastle upon Tyne*	*2012*			
18	*On the Compatibility of Lexical Resources for NooJ*	*Stanković, R., Utvić, M., Vitas, D., Krstev, C., Obradović, I.*	*Automatic Processing of Various Levels of Linguistic Phenomena: Selected Papers from the NooJ 2011 International Conference*		*Cambridge Scholars Publishing*	*Newcastle upon Tyne*	*2012*	*95-107*		
19	*Beszédkorpusz tervezése magyar nyelvű, rejtett Markov-modell alapú szövegfelolvasóhoz (Speech corpus design for Hungarian hidden Markov model based speech synthesis*	*Tóth B., Németh G., Olaszy G*	*Beszédkutatás*	*20*	*MTA Nyelvtudományi Intézet*	*Budapest*	*2012*	*278-295*		

20	*Optimizing HMM Speech Synthesis for Low Resource Devices*	*Tóth, B., Németh G*	*Journal of Advanced Computational Intelligence & Intelligent Informatics*	*16/2*			*2012*	*327-334*		
21	*Recognition and normalization of some classes of named entities in Serbian*	*Krstev, C., Jaćimović, J., Vitas, D.*	*BCI '12 Proceedings of the Fifth Balkan Conference in Informatic*		*ACM*	*New York*	*2012*	*52-57*		
22	*Új módszer nagyméretű beszédadatbázisok formánsadatokkal történő ellátására*	*Abari Kálmán – Olaszy Gábor*	*Beszéd, adatbázis, kutatások*		*Akadémiai Kiadó*	*Budapest*	*2012*	*216-232*		
23	*Bulgarian X-language Parallel Corpus*	*Koeva, S.; Stoyanova, I.; Dekova, R.; Rizov, B.; Genov*	*Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12*		*ELREC*	*Istambul*	*2012*	*51-62*		
24	*The Bulgarian National Corpus: Theory and Practice in Corpus Design*	*Koeva, S.; Stoyanova, I.; Leseva, S.; Dimitrova, T.; Dekova, R.; Tarpomanova, E.*	*Journal of Language Modelling*	*1*			*2012*	*65-110*		

25	*Construction and Exploitation of X-Serbian Bitexts*	*Vitas, D., Krstev, C.*	*Multilingual Processing in Eastern and Southern EU Languages: Low-Resourced Technologies and Translation*		*Cambridge Scholars Publishing*	*Newcastle upon Tyne*	*2012*	*207-227*		
26	*The Polish Sejm Corpus*	*Ogrodniczuk, M.*	*Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12*		*ELRA*	*Istambul*	*2012*	*2219-2223*		
27	*An Approach to Development of Bilingual Lexical Resource*	*Obradović, I., Stanković, R*	*Proceedings of the Fifth Balkan Conference in Informatics BCI 2012, Workshop on Computational Linguistics and Natural Language Processing of Balkan Languages , CLoBL*				*2012*	*101-104*		
28	*Desať rokov národného korpusu/ Ten Years Of the Slovak National*	*Šimková, M.*	*The News Of SAV*	*9*			*2012*	*12-14*		

	Corpus. Správy SAV									
29	*Традиции и новаторство в корпусната лингвистика: Българският национален корпус (Tradition and Innovation in Corpus Linguistics: the Case of the Bulgarian National Corpus).*	*Koeva, S.; Stoyanova, I.; Dimitrova, T.; Leseva, S*	*Journal of the Bulgarian Academy of Sciences*	*3*			*2012*			
30	*Представяне на европейската мрежа META-NET и проекта CESAR в София / The META-NET and CESAR Road Show in Sofia: An Overview*	*Tarpomanova, E*	*Съпоставително езикознание / Sapostavitelno ezikoznanie / Contrastive Linguistics,*	*4*			*2012*			
31	*A System for Named Entity Recognition Based on Local Grammars*	*Cvetana K., Obradović, I., Utvić, M., Vitas, D.*	*Journal of Logic and Computation.*		*Oxford University Press*		*2013*			*yes*

32	*Precíziós, párhuzamos, magyar beszédadatbázis fejlesztése és szolgáltatásai.*	*Olaszy, G.*	*Beszédkutatás*		*MTA Nyelvtudományi Intézet*		*2013*			
33	*Rozvoj jazykových technológií a zdrojov na Slovensku a vo svete (10 rokov Slovenského národného korpusu) / Development of the Human Language Technologies and Resources in Slovakia and in the world (10 years of the Slovak National Corpus)*	*Šimková, M., Garabík, R., Gajdošová, K., Váradi, T., Rehm, G*	*Jazykovedné štúdie*	*31*			*2013*			

A2: LIST OF DISSEMINATION ACTIVITIES								
NO.	Type of activities	Main leader	Title	Date	Place	Type of audience	Size of audience	Countries addressed
1	Conference		FASSBL2010	4–6 Oct, 2010	Dubrovnik, Croatia	LR&T and NLP researcher community in South Slavic and Balkan countries	35	international
2	Conference		FLaReNet 2011 Forum	26–27 May, 2011	Venice, Italy	LR&T and NLP researchers, SMEs and large companies in LR&T and NLP field		international
3	Conference		NooJ2011	13–15 June, 2011	Dubrovnik, Croatia	LR&T and NLP researcher community of developers and users of NooJ NLP processing tool	40	international
4	Conference		InterSpeech 2011	28–31 August, 2011	Firenze, Italy	speech and linguistic research community	over 1000	international
5	Conference		SlaviCorp2011	12–14 September, 2011	Dubrovnik, Croatia	LR&T and NLP researcher community in Slavic countries	40	international
6	Conference		TM-Europe	29–30 September 2011	Warsaw, Poland	Translation professionals, localization software companies	150+	international
7	Conference		LTC2011	25–27 November, 2011	Poznań, Poland	LR&T and NLP researchers, SMEs and large companies in LR&T and NLP field	100s of participa nts	international
8	Invited lecture		Development of Corpus Linguistics in Slovakia	11/26/2011	Faculty of Education, Commenius University, Bratislava, Slovakia	linguistic research and pedagogical community	50	international
9	Conference		Meeting of Slovak	14–16	Častá-	linguistic research	50	international

			Linguists	*November, 2011*	*Papiernička, Slovakia*	*comunity*		
10	*Invited lecture*		*Corpus Linguistics in Slovakia*	*4/14/2011*	*Institut für Slawistik, Universität Wien, Vienna, Austria*	*linguistic research comunity*	*70*	*international*
11			*Researcher's Night*	*9/23/2011*	*Bratislava, Slovakia*	*general public*	*500*	*international*
12			*Researcher's Night*	*9/23/2011*	*Budapest, Hungary*	*general public*		*international*
13	*Conference*		*Conference on Hungarian Computational Linguistics*	*1-2 December 2011*	*Szeged, Hungary*	*Hungarian NLP research community*	*100*	*international*
14	*Invited lecture*		*10th National Conference "New Technologies and standards: digitization of national heritage"*	*22-23 September 2011*	*Belgrade, Serbia*	*computing, library and cultural institution community*	*70*	*international*
15	*Introductory speech*		*41st International Slavistic Conference*	*2-16 September 2011*	*Belgrade, Serbia*	*slavists*	*30*	*international*
16	*Annual meeting for long-life learning*		*Annual Meeting of professors of Computing*	*13-15 January 2012*	*Kragujevac, Serbia*	*professors of mathematics and computing in secondary schools*	*600*	*international*
17	*Annual meeting for long-life learning*		*Annual Meeting of professors of Serbian*	*12-15 January 2012*	*Belgrade, Serbia*	*professors of Serbian in elementary and secondary schools*	*850*	*international*
18	*Open seminar*		*Natural Language Processing Seminar*	*9th January 2012*	*Warsaw, Poland*	*NLP researchers*	*40*	*international*
19	*Workshop*		*Workshop for Master and PhD students on Language Resources and their Processing*	*21st January 2012*	*Belgrade, Serbia*	*master and PhD students in Computing, Library and Information Science, Serbian, and*	*30*	*international*

						Linguistics		
20	Presentation		International Translation Management Conference	29th October 2011	Warsaw, Poland	Localization industry repesentatives, LSP companies, Translation professionals	150+ participants	international
21	Employee meeting		Presentation of the project at ICS PAS	3/9/2011	Warsaw, Poland	IT scientists, linguists	70 employees	international
22	Information portal		Success Stories portal	6/6/2011	Poland	General public	Unknown – available in the Internet	international
23	Presentation		META-FORUM 2011	27-28 June 2011	Budapest, Hungary		ca 200	international
24	Meeting		INFSO.E1 consultation meeting	6-7 December 2011	Paris, France	EC representatives, LT/MT experts	30 participants	international
25	Seminar		NLP Seminar, University of Warsaw	1/3/2012	Warsaw, Poland	IT students	30 participants	international
26	Workshop		Joint workshop of Ľ. Štúr Institute of Linguistics and Institute of Informatics	8 – 10 February 2012	Senec, Slovakia	NLP researchers	20 participants	international
27	Conference		InFuture2011	11/9/2011	Zagreb, Croatia	IT scientists, linguists, ICT companies	cca 100	international
28	Conference		Annual meeting of Croatian Applied Linguistic Society (HDPL2011)	12-14 May 2011	Osijek, Croatia	linguists, IT scientists	cca 150	international
29	Poster presentation		European Day of Languages	9/26/2012	Bratislava, Slovakia	general public	cca 300	international
30	Flyer distribution		Researcher's Night 2012	9/28/2012	Bratislava, Slovakia	general public	cca 400	international
31	Presentations		Open Days 2012	11/8/2012	Bratislava, Slovakia	general public	17	international
32	Presentation,		Presentation of the	11/21/2012	Sofia,	students, teachers	28	international

	workshop		*project at Sofia University St. Kliment Ohridski*		*Bulgaria*			
33	*Conference*		*Young Linguists Meeting 2012*	*5/12/2012*	*Nitra, Slovakia*	*linguists, students, teachers*	*cca 65*	*international*
34	*Conference, roadshow*		*Development of the Human Language Technologies and Resources in Slovakia and in the world (10 years of the Slovak National Corpus)*	*7-8 June, 2012*	*Bratislava, Slovakia*	*linguists, IT scientists, ICT companies, general public*	*cca 100*	*international*
35	*oral paper/presentation*	*FFZG*	*INFuture2011*	*9–11 November, 2011*	*Zagreb, Croatia*	*linguists, IT scientists*	*-*	*Croatia*
36	*oral paper/presentation*	*UBG*	*Crosslingual Language Technology in service of an integrated multilingual Europe - 20 years on-presentation: "Construction and exploitation of X-Serbian bitexts"*	*4–5 May, 2012*	*Hamburg, Germany*	*linguists, IT scientists, ICT companies, general public*	*50*	*international*
37	*oral paper/presentation*	*UBG*	*14th International Scientific Conference of the Commission for Word Derivation of the Interantional Slavic Committee*	*28–30 May, 2012*	*Belgrade, Serbia*	*linguists*	*60*	*international*
38	*flayer distribution*	*UBG*	*BCI '12 - the Fifth Balkan Conference in Informatics*		*Novi Sad, Serbia*	*linguists, IT scientists, ICT companies, general public*	*80*	*international*
39	*complex event (road show)*	*IBL*	*CESAR / META-NET Roadshow, Sofia 2012*	*5/2/2012*	*Sheraton Sofia Hotel Balkan, Bulgaria*	*linguists,stakeholders, politicians, IT scientists, ICT companies, general public*	*120*	*international*

40	*poster presentation*	*IPUP*	*International NooJ Conference 2012*	*14–16 June, 2012*	*Paris, France*	*Linguists, IT scientists*	*50*	*international*
41	*poster presentation*	*IPUP*	*International NooJ Conference 2012*	*14–16 June, 2012*	*Paris, France*	*Linguists, IT scientists*	*50*	*international*
42	*oral paper/presentation*	*IPUP*	*ICIST 2011 - Second International Conference on Internet Society Technology and Management*	*6–9 March, 2012*	*Kopaonik, Serbia*	*Linguists, IT scientists*	*60*	*international*
43	*demo*	*ULODZ*	*Poznań Linguistic Meeting*	*7–10 September, 2012*	*Poznań, Poland*	*Linguists, IT scientists*	*50*	*Polish*
44	*complex event (road show)*	*IBL*	*CESAR / META-NET Roadshow 2012 (Sofia, Bulgaria)*	*5/2/2012*	*Hotel Sheraton - Sofia, Bulgaria*	*linguists,stakeholders, politicians, IT scientists, ICT companies, general public*	*120*	*international*
45	*oral paper/presentation*	*IBL*	*CESAR Road show in Serbia, October 29th, 2012*	*10/29/2012*	*Hotel Hyatt Regency - Belgrade, Serbia*	*linguists,stakeholders, politicians, IT scientists, ICT companies, general public*	*120*	*international*
46	*oral paper/presentation*	*IBL*	*Eight International Conference on Language Resources and Evaluation (LREC'12)*	*21–27 May, 2012*	*Istanbul, Turkey*	*linguists, IT scientists, ICT companies, general public*	*1000*	*international*
47	*flyer distribution*	*IBL*	*70th Anniversary of the Bulgarian Academic Lexicography: Sixth National Conference (with Internation Participation) on Lexicography and Lexicology*	*24–25 October, 2012*	*Sofia, Bulgaria*	*linguists, IT scientists, general public*	*30*	*international*
48	*complex event (road show)*	*IBL*	*META-FORUM 2012*	*19–21 June, 2012*	*Brussels, Belgium*	*linguists,stakeholders, politicians, IT scientists, ICT*	*300*	*international*

						companies		
50	*flyer distribution*	*IBL*	*Policies on minority and immigrant languages in Bulgaria / Политики за малцинствените и имигрантски езици в България*	*11/9/2012*	*British Council, Sofia, Bulgaria*	*languages, stakeholders, politicians*	*30*	*Bulgaria*
51	*flyer distribution*	*IBL*	*Overview of Language Rich Europe project*	*9/26/2012*	*Sheraton Hotel, Sofia, Bulgaria*	*languages, stakeholders, politicians*	*50*	*international*
52	*oral paper/presentation*	*IBL*	*Scientific Pre-Christmas Party*	*12/20/2012*	*Sofia, Institute for Bulgarian Language*	*linguists*	*40*	*Bulgaria*
53	*flyer distribution*	*IBL*	*Scientific Pre-Christmas Party*	*12/20/2012*	*Sofia, Institute for Bulgarian Language*	*linguists*	*40*	*Bulgaria*
54	*radio broadcast*	*IBL*	*The META-NET White Paper Series were presented in the radio stations in Bulgaria. 3 radio broadcasts - Bulgarian National Radio / Radio Sofia, Radio Vitosha, Radio Veselina.*		*Bulgaria*	*general public*	*-*	*Bulgaria*
55	*TV broadcast*	*IBL*	*Information about META-NET White Paper Series was aired on Bulgarian TV (4 TV channels - BTV, TV Alpha, TV7, TV Evropa)*		*Bulgaria*	*general public*	*-*	*Bulgaria*
57	*poster presentation*	*LSIL*	*European Day of Languages*	*9/26/2012*	*Bratislava, Slovakia*	*general public, linguists, students*	*50*	*Slovakia*
58	*flyer distribution, poster*	*LSIL*	*Researcher's Night 2012*	*9/28/2012*	*Bratislava, Slovakia*	*general public, linguists, students*	*40*	*Slovakia*

	presentation							
59	*presentations*	*LSIL*	*Open Days 2012*	*11/8/2012*	*Bratislava, Slovakia*	*general public, linguists, students*	*50*	*Slovakia*
60	*presentation, workshop*	*LSIL*	*Presentation of the project at Sofia University St. Kliment Ohridski*	*21/11/2012*	*Sofia, Bulgaria*	*general public, linguists, students*	*40*	*Bulgaria*
61	*presentations*	*LSIL*	*Young Linguists Meeting 2012*	*5–7 December, 2012*	*Nitra, Slovakia*	*linguists, students*	*60*	*Slovakia*
62	*presentations, flyer distribution, road-show*	*LSIL*	*Development of the Human Language Technologies and Resources in Slovakia and in the world (10 years of the Slovak National Corpus)*	*7–8 June, 2012*	*Bratislava, Slovakia*	*linguists,stakeholders, politicians, IT scientists, ICT companies, general public*	*120*	*international*
63	*Invited talk at the International Translation Management Conference 2012*	*IPIPAN/ULODZ*	*International Translation Management Conference 2012*	*4–6 October, 2012*	*Warsaw, Poland*	*linguists, MT specialists, IT specialists*	*60*	*international*
64	*Paper/poster presentation at LREC2012*	*ULODZ/IPIPAN*	*LREC 2012 Conference*	*21–27 May 2012*	*Istanbul, Turkey*	*linguists,stakeholders, politicians, IT scientists, ICT companies, general public*	*1000*	*international*
67	*oral paper/presentation*	*UBG*	*SlaviCorp 2011 : SlaviCorp - Corpora of Slavic Languages; presentation: Named Entities in the Corpus of Contemporary Serbian*	*12–14 September, 2011*	*Dubrovnik, Croatia*	*LT specialists, linguists*	*60*	*international*
68	*oral paper/presentation*	*UBG*	*SlaviCorp 2011 : SlaviCorp - Corpora of Slavic Languages; presentation: Corpus of*	*12–14 September, 2011*	*Dubrovnik, Croatia*	*LT specialists, linguists*	*60*	*international*

			Contemporary Serbian - Annotation Strategies					
69	*oral paper/presentation*	*UBG*	*The Annual Day of the Faculty of Mathematics, December 21st, 2012.*	*12/21/2012*	*Belgrade, Serbia*	*LT specialists, linguists*	*70*	*Serbia*
70	*oral paper/presentation*	*UBG*	*8th FASSBL conference - Formal Aspects of South Slavic and Balkan Languages; presentation: Derivational Patterns in E-Dictionaries of Serbian*	*9/19/2012*	*Dubrovnik, Croatia*	*LT specialists, linguists*	*50*	*international*
71	*oral paper/presentation*	*UBG*	*10th National Conference "New Technologies and standards: digitization of national heritage; presentation: Serbian Language and its Resources*	*9/22/2011*	*Belgrade, Serbia*	*LT specialists, linguists*	*50*	*international*
72	*oral paper/presentation*	*UBG*	*41st International Slavistic Conference; presentation "Language Resources"*	*9/2/2011*	*Belgrade, Serbia*	*linguists*	*60*	*international*
73	*oral paper/presentation*	*UBG*	*Open Standards in Science - On the occasion of 10 years of Creative Commons in Serbia; presentation: Language Resources in META-SHARE*	*12/13/2012*	*Belgrade, Serbia*	*LT specialists, linguists*	*80*	*international*
74	*oral paper/presentation*	*UBG*	*18th Assembly of the Serbian Association of Academic Libraries; presentation "Serbian Language and its Resources"*	*6/5/2012*	*Belgrade, Serbia*	*LT specialists, linguists*	*100*	*international*

75	*conference, road-show*	*HASRIL*	*The Hungarian Language in the Digital Age*	*18/01/2013*	*Budapest, Hungary*	*linguists,stakeholders, politicians, IT scientists, ICT companies, general public*	*120*	*international*

Report on societal implications

Replies to the following questions will assist the Commission to obtain statistics and indicators on societal and socio-economic issues addressed by projects. The questions are arranged in a number of key themes. As well as producing certain statistics, the replies will also help identify those projects that have shown a real engagement with wider societal issues, and thereby identify interesting approaches to these issues and best practices. The replies for individual projects will not be made public.

A General Information *(completed automatically when **Grant Agreement number** is entered.*

Grant Agreement Number:	**271022**
Title of Project:	**Central and South-East European Resources**
Name and Title of Coordinator:	Tamás Váradi

B Ethics

1. Did your project undergo an Ethics Review (and/or Screening)? • If Yes: have you described the progress of compliance with the relevant Ethics Review/Screening Requirements in the frame of the periodic/final project reports? Special Reminder: the progress of compliance with the Ethics Review/Screening Requirements should be described in the Period/Final Project Reports under the Section 3.2.2 *'Work Progress and Achievements'*	***No***
2. Please indicate whether your project involved any of the following issues (tick box) :	***NO***
RESEARCH ON HUMANS	
• Did the project involve children?	NO
• Did the project involve patients?	NO
• Did the project involve persons not able to give consent?	NO
• Did the project involve adult healthy volunteers?	NO
• Did the project involve Human genetic material?	NO
• Did the project involve Human biological samples?	NO
• Did the project involve Human data collection?	NO
RESEARCH ON HUMAN EMBRYO/FOETUS	
• Did the project involve Human Embryos?	NO
• Did the project involve Human Foetal Tissue / Cells?	NO
• Did the project involve Human Embryonic Stem Cells (hESCs)?	NO
• Did the project on human Embryonic Stem Cells involve cells in culture?	NO
• Did the project on human Embryonic Stem Cells involve the derivation of cells from Embryos?	NO
PRIVACY	
• Did the project involve processing of genetic information or personal data (eg. health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction)?	NO
• Did the project involve tracking the location or observation of people?	NO
RESEARCH ON ANIMALS	
• Did the project involve research on animals?	NO
• Were those animals transgenic small laboratory animals?	NO
• Were those animals transgenic farm animals?	NO
• Were those animals cloned farm animals?	NO
• Were those animals non-human primates?	NO
RESEARCH INVOLVING DEVELOPING COUNTRIES	
• Did the project involve the use of local resources (genetic, animal, plant etc)?	NO

• Was the project of benefit to local community (capacity building, access to healthcare, education etc)?	NO
DUAL USE	
• Research having direct military use	NO
• Research having the potential for terrorist abuse	NO

C Workforce Statistics

3. Workforce statistics for the project: Please indicate in the table below the number of people who worked on the project (on a headcount basis).

Type of Position	**Number of Women**	**Number of Men**
Scientific Coordinator	0	1
Work package leaders	4	1
Experienced researchers (i.e. PhD holders)		
PhD Students		
Other		

4. How many additional researchers (in companies and universities) were recruited specifically for this project?	**0**
Of which, indicate the number of men:	

D Gender Aspects

5. Did you carry out specific Gender Equality Actions under the project?	○	NO

6. Which of the following actions did you carry out and how effective were they?

		Not at all effective				Very effective
❑	Design and implement an equal opportunity policy	○	○	○	○	○
❑	Set targets to achieve a gender balance in the workforce	○	○	○	○	○
❑	Organise conferences and workshops on gender	○	○	○	○	○
❑	Actions to improve work-life balance	○	○	○	○	○

○ Other:

7. Was there a gender dimension associated with the research content – i.e. wherever people were the focus of the research as, for example, consumers, users, patients or in trials, was the issue of gender considered and addressed?

○ Yes- please specify

X No

E Synergies with Science Education

8. Did your project involve working with students and/or school pupils (e.g. open days, participation in science festivals and events, prizes/competitions or joint projects)?

X Yes- please specify: in some tasks, students weve involved, eg. digitalization and other manual works

○ No

9. Did the project generate any science education material (e.g. kits, websites, explanatory booklets, DVDs)?

X Yes- please specify: video lectures on NooJ

○ No

F Interdisciplinarity

10. Which disciplines (see list below) are involved in your project?

X Main discipline[2]: 1.1, 5.4, 6.2

○ Associated discipline[2]: ○ Associated discipline[2]:

G Engaging with Civil society and policy makers

11a Did your project engage with societal actors beyond the research community? *(if 'No', go to Question 14)*	○	No

11b If yes, did you engage with citizens (citizens' panels / juries) or organised civil society (NGOs, patients' groups etc.)?

○ No

○ Yes- in determining what research should be performed

○ Yes - in implementing the research

○ Yes, in communicating /disseminating / using the results of the project

11c In doing so, did your project involve actors whose role is mainly to organise the dialogue with citizens and organised civil society (e.g. professional mediator; communication company, science museums)?	○ ○	Yes No

12. Did you engage with government / public bodies or policy makers (including international organisations)

○ No

○ Yes- in framing the research agenda

[2] Insert number from list below (Frascati Manual).

○ Yes - in implementing the research agenda

X Yes, in communicating /disseminating / using the results of the project

13a Will the project generate outputs (expertise or scientific advice) which could be used by policy makers?

○ Yes – as a **primary** objective (please indicate areas below- multiple answers possible)

X Yes – as a **secondary** objective (please indicate areas below - multiple answer possible)

○ No

13b If Yes, in which fields?

Agriculture		Energy		Human rights	
Audiovisual and Media		Enlargement	X	**Information Society**	
Budget		Enterprise		Institutional affairs	
Competition		Environment		Internal Market	
Consumers		External Relations		Justice, freedom and security	
Culture		External Trade		Public Health	
Customs		Fisheries and Maritime Affairs		Regional Policy	
Development Economic and		Food Safety		Research and Innovation	
Monetary Affairs		Foreign and Security Policy		Space	
Education, Training, Youth		Fraud		Taxation	
Employment and Social Affairs		Humanitarian aid		Transport	

13c If Yes, at which level?

○ Local / regional levels
X National level
○ European level
○ International level

H Use and dissemination

14. How many Articles were published/accepted for publication in peer-reviewed journals?	**33**
To how many of these is open access[3] provided?	
How many of these are published in open access journals?	
How many of these are published in open repositories?	
To how many of these is open access not provided?	
Please check all applicable reasons for not providing open access:	
□ publisher's licensing agreement would not permit publishing in a repository □ no suitable repository available □ no suitable open access journal available □ no funds available to publish in an open access journal □ lack of time and resources □ lack of information on open access □ other[4]: ……………	
15. How many new patent applications ('priority filings') have been made? *("Technologically unique": multiple applications for the same invention in different jurisdictions should be counted as just one application of grant).*	**0**

16. Indicate how many of the following Intellectual Property Rights were applied for (give number in each box).	Trademark	**0**
	Registered design	**0**
	Other	**0**

17. How many spin-off companies were created / are planned as a direct result of the project?	**0**
Indicate the approximate number of additional jobs in these companies:	**0**

18. Please indicate whether your project has a potential impact on employment, in comparison with the situation before your project:

□	Increase in employment, or	□	In small & medium-sized enterprises
□	Safeguard employment, or	□	In large companies
□	Decrease in employment,	X	None of the above / not relevant to the project
□	Difficult to estimate / not possible to quantify		

19. For your project partnership please estimate the employment effect resulting directly from your participation in Full Time Equivalent (*FTE = one person working fulltime for a year*) jobs:	*Indicate figure:*
Difficult to estimate / not possible to quantify	□

[3] Open Access is defined as free of charge access for anyone via Internet.
[4] For instance: classification for security project.

I Media and Communication to the general public

20. As part of the project, were any of the beneficiaries professionals in communication or media relations?

X Yes ○ No

21. As part of the project, have any beneficiaries received professional media / communication training / advice to improve communication with the general public?

X Yes ○ No

22 Which of the following have been used to communicate information about your project to the general public, or have resulted from your project?

□ Press Release
□ Media briefing
□ TV coverage / report
X Radio coverage / report
X Brochures /posters / flyers
X DVD /Film /Multimedia

□ Coverage in specialist press
□ Coverage in general (non-specialist) press
□ Coverage in national press
□ Coverage in international press
X Website for the general public / internet
X Event targeting general public (festival, conference, exhibition, science café)

23 In which languages are the information products for the general public produced?

□ Language of the coordinator
X Other language(s)

X English

Question F-10: Classification of Scientific Disciplines according to the Frascati Manual 2002 (Proposed Standard Practice for Surveys on Research and Experimental Development, OECD 2002):

FIELDS OF SCIENCE AND TECHNOLOGY

1. NATURAL SCIENCES

1.1 Mathematics and computer sciences [mathematics and other allied fields: computer sciences and other allied subjects (software development only; hardware development should be classified in the engineering fields)]
1.2 Physical sciences (astronomy and space sciences, physics and other allied subjects)
1.3 Chemical sciences (chemistry, other allied subjects)
1.4 Earth and related environmental sciences (geology, geophysics, mineralogy, physical geography and other geosciences, meteorology and other atmospheric sciences including climatic research, oceanography, vulcanology, palaeoecology, other allied sciences)
1.5 Biological sciences (biology, botany, bacteriology, microbiology, zoology, entomology, genetics, biochemistry, biophysics, other allied sciences, excluding clinical and veterinary sciences)

2 ENGINEERING AND TECHNOLOGY

2.1 Civil engineering (architecture engineering, building science and engineering, construction engineering, municipal and structural engineering and other allied subjects)
2.2 Electrical engineering, electronics [electrical engineering, electronics, communication engineering and systems, computer engineering (hardware only) and other allied subjects]
2.3. Other engineering sciences (such as chemical, aeronautical and space, mechanical, metallurgical and materials engineering, and their specialised subdivisions; forest products; applied sciences such as geodesy, industrial chemistry, etc.; the science and technology of food production; specialised technologies of interdisciplinary fields, e.g. systems analysis, metallurgy, mining, textile technology and other applied subjects)

3. MEDICAL SCIENCES

3.1 Basic medicine (anatomy, cytology, physiology, genetics, pharmacy, pharmacology, toxicology, immunology and immunohaematology, clinical chemistry, clinical microbiology, pathology)
3.2 Clinical medicine (anaesthesiology, paediatrics, obstetrics and gynaecology, internal medicine, surgery, dentistry, neurology, psychiatry, radiology, therapeutics, otorhinolaryngology, ophthalmology)
3.3 Health sciences (public health services, social medicine, hygiene, nursing, epidemiology)

4. AGRICULTURAL SCIENCES

4.1 Agriculture, forestry, fisheries and allied sciences (agronomy, animal husbandry, fisheries, forestry, horticulture, other allied subjects)
4.2 Veterinary medicine

5. SOCIAL SCIENCES
5.1 Psychology
5.2 Economics
5.3 Educational sciences (education and training and other allied subjects)
5.4 Other social sciences [anthropology (social and cultural) and ethnology, demography, geography (human, economic and social), town and country planning, management, law, linguistics, political sciences, sociology, organisation and methods, miscellaneous social sciences and interdisciplinary , methodological and historical S1T activities relating to subjects in this group. Physical anthropology, physical geography and psychophysiology should normally be classified with the natural sciences].

6. HUMANITIES
6.1 History (history, prehistory and history, together with auxiliary historical disciplines such as archaeology, numismatics, palaeography, genealogy, etc.)
6.2 Languages and literature (ancient and modern)
6.3 Other humanities [philosophy (including the history of science and technology) arts, history of art, art criticism, painting, sculpture, musicology, dramatic art excluding artistic "research" of any kind, religion, theology, other fields and subjects pertaining to the humanities, methodological, historical and other S1T activities relating to the subjects in this group]