

# PROJECT OBJECTIVES



## TABLE OF CONTENTS

<b>EU dimension</b>	<b>2</b>
<b>Maturity of the technical solution</b>	<b>2</b>
<b>Target outcome and expected impact</b>	<b>3</b>
<b>Long term viability</b>	<b>5</b>
<b>META-SHARE nodes</b>	<b>6</b>
<b>Work progress and achievements during the period</b>	<b>7</b>
Work progress and achievements by WPs	7
WP1 Management (HASRIL)	7
WP2 Analysis and selection of language resources (IBL)	8
WP3: Enhancing language resources (IPIPAN)	11
WP4 Cross-national collaboration and Pilot service (BME-TMIT)	20
WP5 Outreach, awareness and sustainability (FFZG)	21

The main goals of CESAR project are:

- 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 concurrent 6.1 pilot projects and the META-NET network of excellence - and where useful with other relevant multi-national forums or activities, such as FlaReNET and CLARIN - to ensure consistent approaches, practices and standards aimed at ensuring a wider accessibility of 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;<sup>1</sup>
- 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 by stimulating ICT-based cross-lingual communication, collaboration and participation.<sup>2</sup>

<sup>1</sup> As defined in the Call

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 these 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 at the modern digital market: everyday end-users, professional end-users (business, administration, media, education, libraries, etc.) as well as expertise holders (researchers, industrialists, policy makers, etc). The concern of the project is a careful investigation of the needs of various types of users – from individual users to large multinational organization's – 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.

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

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

## EU dimension

The CESAR project is fully committed to providing synergy with relevant strategies and activities at the European, national and regional level. The partners are key players in the language technology scene for their respective languages and have built extensive links with LRT centres across the whole of Europe but especially in the geographical region that the consortium covers. At the same time, they have served and will continue to serve as catalytic force in the language technology scene of their countries. Typically, they have led national consortia involving the other major partners in the countries and have an established track record of their capability to mobilise their national partners. All partners are members of the CLARIN infrastructure project. As such, they are actively involved in building links with national funders, staging nationwide events raising awareness of the potential importance of language technology and language resources. The widespread professional and official links developed in the past two years as a result of their work in CLARIN can be usefully deployed in similar activities in the interest of the CESAR project.

## 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 have emerged in the respective language communities autonomously, whereas the current state-of-the-art already would require and allow for improving over such baseline situation, necessitating that existing partial solutions are now going to be combined, tested, and integrated. In each country in the past, various research communities in academia and industry have started compiling either general-purpose corpora, or, at

---

<sup>2</sup> As it was underlined in the Competitiveness and Innovation Framework Programme (CIP) ([ftp.cordis.europa.eu/pub/fp7/ict/docs/language-technologies/excerpts-on-theme-6-multilingual-web\\_en.pdf](ftp.cordis.europa.eu/pub/fp7/ict/docs/language-technologies/excerpts-on-theme-6-multilingual-web_en.pdf))

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. Likewise, the syntactic-semantic markup of these partially thematically overlapping resources tends to be similar, but not in sufficient detail to make these resources and tools interoperable: they often contain idiosyncrasies, due to special purposes they were designed for, or because of language-dependent phenomena. Since until recently there have been no complex large scale applications either at the national levels, or across borders or language communities, many of the resources available are not yet in the state of being ready for straightforward integration. Likewise, most of the tools have been developed and tested on single corpora even if they allow to be applied to others, and to be ported to other languages. It needs thorough investigation to what extent the materials and tools existing in and applicable to the separate languages correspond to each other, where and by what means they can be made optimally extendable so that a balanced range of resources is established for Central and South European languages that are harmonized with each other.

The resources in this way allow for and boost generation of statistical models generally required in linguistic analysis and synthesis pipelines. Methodological considerations used in models existing for a certain language are going to be applied and ported to new languages. The resources are made available as reusable basic building blocks for further, more advanced applications, such as information retrieval, document classification, query expansion, synonymy search, information extraction. Multilinguality-specific tasks (CLIR, MAT, MT, etc.) can in turn draw on these intermediary applications.

The CESAR project tries to bring 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 make them reachable 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 CESAR project outcomes. The project partners not just predict 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 not just following the predictions but is taking part in paving the way for these languages to the European single digital market.

Given the amount of time and investment of joint effort required to accomplish this, it is necessary to carry out such activities in a concerted way, which also guarantee the implementation of consistent approaches and practices. Good documentation and manuals are individually available for each of the resources, but they need to be harmonized as well in order to document, link, and upgrade them according to agreed standards and guidelines, where the partners can build on results of previous joint projects such as MULTEXT-EAST and TELRI.

Creating thoroughly descriptive metadata based on the current metadata allow for cross-alignment and better matching between tools and corpora, which ensures that both are exploited to the maximum, without unnecessary duplication of level of knowledge-rich, interoperable resources

### **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 creates, populates and operates 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 of CESAR were be 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 is 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 is 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 project(s), but perspective for a longer time.

Next to achieving the project milestones, the partners intend to assess intermediate and final results according to the impact on national LRT communities, in terms of the following: appearance of LRT in national priorities scientific programs, number of meetings of the national LRT society, number of nodes in network of centres offering resources and number of "transactions" between them, indicating temporal and geographical outreach, links established between certain types of tools as nodes etc.

## 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 will be made available either via servers of the original creators and/or 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 FLReNET 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 should be constantly improved to support user requirements.
- *Documentation*: Based on resources documentation, guidelines should be 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 should 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.

Above-mentioned aspects have been discussed by the consortium members in order to indicate clear-cut criteria for measuring project sustainability, driven by the following considerations:

*Availability*: The project tools, resources and documentation must be incessantly available, with negligible level of inaccessibility.

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, and it may be expected that the national and EU funding could be available for maintaining these resources. Such large-scale pre-competitive resources have been only marginally funded by industrial partners. Several scenarios of maintenance are foreseen after the CESAR project ends:

- 1) national or EU funding for maintenance of resources submitted in the CESAR project;
- 2) national or EU funding for inclusion and adaptation of resources to wider networks, repositories, infrastructures;
- 3) commercial funding of maintenance of resource(s) from single source as competitive advantage;
- 4) commercial funding of maintenance of resource(s) from several sources as pre-competitive cooperation;
- 5) any combination of the previous scenarios.

With the next coming round of development of language resources and tools and by making CESAR resources more accessible, we expect that scenarios 3, 4 and 5 will dominate over 1 and 2.

This will be one of keys to sustainability of produced resources and tools.

While the maintenance could be considered a lower-cost activity, development and adaptation of resources to user needs is supposed to be more demanding. In this respect within the CESAR project four main directions of activities were stressed:

- 1) enhancing resources and tools in size, coverage, precision, recall, accuracy etc.;
- 2) adapting new resources and tools to become compliant with the agreed standard for interoperability;
- 3) upgrading resources and tools by combining them with other resources and tools (e.g. pipelining) in order to achieve desired level of interoperability;
- 4) adapting user-interfaces to fulfill user requirements.

The user requirements will be collected, investigated, planned and provided in collaboration with our possible largest users (e.g. META-NET and similar initiatives).

### **META-SHARE nodes**

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.

CESAR META-SHARE nodes are brand for long-term maintenance for 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 committed 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).

List of CESAR nodes set up at the end of the project:

- 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-share.matf.bg.ac.rs>,
- LSIL node: <https://metashare.korpus.sk/>



## Work progress and achievements during the period

### Work progress and achievements by WPs

#### WP1 Management (HASRIL)

This section is focusing on the main achieved results of WP1. A detailed and exhaustive description of the managerial work will be provided in the separate document ‘Project management and use of resources’.

Main effort of this WP is to effectively coordinate and monitor the project in administrative and ensure that the technical and financial goals of the project were realised successfully. The tasks of WP are continuously maintained through the project lifetime. This effort involves continuous controlling, monitoring and assuring the timeliness and quality of the deliverables, keeping adequate records of the progress of the project, scheduling and organising regular meetings, reporting to the EC and relaying their views to the consortium, managing risks and establishing quality standards for the whole project.

WP1 required daily routines both from project coordinator and all partners, but coordinator served as the managing link between EC, other projects (T4ME, META-NORD, METANET4U, META-NET and infrastructures CLARIN and FLARENET). In technical and particular issues WP and group leaders were communicating the interests of the project.

#### WP1.1 Project coordination and management (HASRIL, BME-TMIT, FFZG, IPIAN, ULodz, UBG, IPUP, IBL, LSIL)

Project was coordinated by HASRIL and WPs by WP leaders. As a part of the management task the internal and external project page was set-up and is maintained by the coordinator ([www.cesar-project.net](http://www.cesar-project.net)). For coordination tasks several on-line facilities were used (eg. Skype, GoogleDoc, GoogleSpreadsheet).

#### WP1.2 Administrative and financial management (HASRIL)

Budgetary and financial practices were coordinated by the project coordinator. He served as an advisor in budgetary issues and acted as controller for partners. All financial practices of the project were in accordance with the EU guidelines. A special effort of financial tasks was made at M18 and M24 (financial report according to EC guidelines).

#### WP1.3 Reporting and submission of deliverables (HASRIL, BME-TMIT, FFZG, IPIAN, ULodz, UBG, IPUP, IBL, LSIL)

The successful maintenance of this task was ensured by the effective cooperation of the partners. Deliverables at M18 were submitted in due time, deliverables in M24 were completed with minor delay and the whole deliverables packages was submitted 8 days late due to delays within META-NET in particular META-SHARE and the accumulated workload of some key partner in the last month of the project.

All relevant information concerning the submitted deliverables (containing the justification of the late submission) can be found in the *Deliverables* and *Milestones* table.

#### WP1.4 Quality Control and Risk Management (HASRIL, BME-TMIT, FFZG, IIPAN, ULodz, UBG, IPUP, IBL, LSIL)

The task includes monitoring all project activities and ensuring the successful completion of work packages, which can be measured by the rate of efficiency of the submitted deliverables. The successful realization of the task was the fruit of each WP leaders.

#### WP1.5 Internal Communication (HASRIL, IPUP)

The project is using several communication channels to ensure the effective communication and work, which in fact leads to the efficient definition and implementation of the communication protocols in term of supporting the interaction within the consortium members. During the project participants were involved in face-to-face meeting, audio and video conferences (Skype) and an internal e-mail list were created ([cesar@nytud.hu](mailto:cesar@nytud.hu)) which is maintained by the coordinator. A detailed schedule of the communication can be found in the list of events (see section 'Project management and use of resources').

#### WP1.6 Collaboration with other projects (HASRIL, BME-TMIT, FFZG, IIPAN, ULodz, UBG, IPUP, IBL, LSIL)

The CESAR project is in close cooperation and collaboration with the projects T4ME, METANET4U and META-NORD(META-NET, META-SHARE). This effective cooperation and collaboration can be seen in several deliverables such as the D4.4 (*Second batch of language resources*), D4.5 (*Third batch of language resources*) D5.1 (*Action plan*) and D5.3 (*Long time sustainability plan*).

In addition to the mentioned above, the consortium is taking an active role and part in META-NET boards and working groups. The cooperation with META-NET was maintained according to the following:

- The coordinator kept regular, often daily contact with network manager Georg Rehm and held regular discussions with the coordinators of METANET4U and METANORD.
- Marko Tadić acted as representative for CESAR within the META-NET Communication Working Group and kept regular contact with its leader John Judge
- Svetla Koeva acted as representative for CESAR within the Language White Paper Working Group
- Maciej Ogrodniczuk and György Szaszák took part in of the Metadata and IPR Working Group
- Tamás Váradi and Adam Przepiórkowski were members of the Executive Board of META-NET
- Piotr Pęzik was member of META-TRUST
- All partners took part in the dissemination effort of META-NET(translations + distribution of META-NET materials)

The CESAR project also developed and maintained collaborative ties with infrastructure projects (e.g. CLARIN-ERIC, FLReNeT) and projects, networks of related interest.

Submitted deliverables:

- D1.2b Periodic report at M18
- D1.2c Periodic report at M24

#### WP2 Analysis and selection of language resources (IBL)

This Work Package aims to chart the national scene and respectively the language community landscape as well as to identify, select, acquire and/or obtain permission to rework and 'publish' own and third parties' language resources and tools. CESAR encompasses a large variety of language resources, including language data, such as written and spoken corpora (annotated or in raw form, monolingual as well as multilingual), lexical and terminological databases, grammars, ontologies, etc.; language processing and annotation tools and technologies. The target users are developers and researchers both in 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, EC institutions, and private investors.

**WP2.1 Charting the national scene – producing „Language whitepaper” (HASRIL, BME-TMIT, FFZG, IPIAN, ULodz, UBG, IPUP, IBL, LSIL)**

Although the focus of this task was on production of the Language Whitepapers for respective languages, several other actions were done (however in connection with the whitepaper series). Within the first months partners managed to prepare the first version of the LWPs (the publication of the first, draft version was scheduled to the META-FORUM 2011 conference – 26th June 2011). After the draft version partners had opportunity to update and complete and also to localize to national languages. The foreseen aim of this work was to publish all LWPs at a well-known publishing house. The LWP series were prepared (and the ones covered by the CESAR participants also submitted as deliverable D2.1) as a coordinated action for the four projects in META-NET and the official version was published in October 2012.

The function of a language report was to recapitulate and document the language community landscape for a given language community by volume and the whole Europe by the whole series. In connection with the LWPs partners identified relevant researchers and projects, policy makers, industry representatives, language communities and additional stakeholders (the elaborated list of relevant stakeholders was continuously updated during the project period).

Within this task the following questions were charted in the involved six countries:

- *Language community*: number of speakers worldwide, number of web pages in that language, other relevant quantitative elements including e.g. estimated volume of translations as source or target language, main trading partners (within and outside the EU), etc.
- *Role of the language in question in the respective country/language community*: legal framework regarding the use of national language(s); institutional communication and local administration; place and function in the media: TV, cinema, press; place and function in the software and digital media (e.g. games) industry e.g. degree of localisation; policies and public programmes in support of language, e.g. language learning, book translation etc.
- *Research community*: estimated size of the research community in the areas of NLP and ST, including specialist groups (e.g. machine translation, information retrieval/extraction); main universities and research centres in NLP and ST; teaching curricula and number of graduates in recent years; national programmes/agencies in support of language technology; main gaps e.g. underdeveloped human or technical resources; activities at national level, their relevance for addressing the identified gaps.
- *Language service industry*: qualitative and quantitative analysis of the local translation, localization and interpretation industries; description and actual/estimated number of businesses and professionals with an indication of leading companies; degree of sophistication and ICT use of the service industry.
- *Language technology industry*: qualitative and quantitative analysis of the local industrial landscape; estimated number of vendors and developers (companies as well as individuals); a description of the main existing LT products and services, and of their actual or potential users (public at large, business/professional users).
- *Policy makers*: politicians, administration, media, funding agencies, affecting the language related community and digital market.
- *Demand side*: role of language-technology products and services within the Internet, digital media and telecommunications sectors, by ways of examples; where applicable: "success stories" i.e. examples of use of language technology by businesses and administrations.
- *Legal provisions*: national intellectual-property and digital-copyright regulations related to language resources i.e. databases and software.
- *Various types of users*: analysis of the needs (of different types of users – from individual users to large multinational organisations (practically all stakeholders at the modern digital market: everyday

end-users, professional end-users (business, administration, media, education, libraries, etc.) as well as expertise holders (researchers, industrialists, policy makers, etc).

- *Contacts information:* (i.e., name, phone number, affiliation, email and postal address) on the international and especially on the national level of representatives of the following stakeholder types: research, politics, administration, funding agencies, LT user industries, LT provider industries, journalists, language communities.

During the work a joint stakeholders contact database from consortium countries (D.2.2) was collected and made available to all consortium partners as well as to other META-NET partners. This database covers individuals (experts), institutions (research, national-funding agencies, government) and companies (producers and important users) dealing with LRT. This database is used primarily for dissemination purposes, but it remains available for other purposes as well. The first version was updated at M18.

## WP2.2 Identification of resources actually or potentially available to the consortium (HASRIL, BME-TMIT, FFZG, IIPAN, ULodz, UBG, IPUP, IBL, LSIL)

In the frame of this task the already developed or under development language resources and tools were identified and a self-constructed model for the description of language resources and tools was adopted.

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 (each of them specified according to different sets of criteria) that were considered representative and indicative for the selection of language resources – D.2.4. The first indicator is general, thus assessing the indicator according to general yes/no criteria. All evaluated resources for a given language are assessed according to a set of criteria such as:

- All selected resources are state-of-the-art representatives of their type for a given language. (yes, no)
- Equally valuable representatives are all included in the selection. (yes, no)
- 16 criteria altogether.

The next two indicators – Total point Value (TPV) 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 notions of availability, quality, quantity and standards are further specified and taken into account in the process of the TPV identification. A technique, supplementing previous approaches, while defining exact measures for quality and quantity aspects and incorporating the standardisation into the quality section, is developed. An important source of data for our analysis are the tables for individual languages produced in the Language White Papers. The Language White Papers provide an overview of the current situation of language technology support. The rating of existing resources and tools is based on educated estimations by several leading experts using the following criteria: Quantity, Availability, Quality, Coverage, Maturity, Sustainability and Adaptability. The fourth indicator (Proportion between the selected resources developed inside and outside the consortium) 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.

Thus one of the main outcomes of this task is the established methodology and criteria that allow partners to assess the quality and importance of language resources and tools.

The partners have contacted research institutions and private companies in their countries that are developers and copyrights owners of language resources and tools seeking the important and relevant information. As a result, partners identified the resources which are or can be made available to them and established a catalogue of written and spoken language resources and tools.

The description of the resources in a well structured and documented format was provided in D.2.3a, D.2.3b and D.2.3c. Thus the next important outcome is the Report on resources (actually or potentially available to the consortium). The report gives an overview of the main language resources of the Central and South-East Europe. It is compiled to give more than 30 types of information on resources of six languages. Tables containing values of the commonly accepted metadata scheme were constructed to gather all important information concerning available and potential language resources. CESAR adopted the metadata scheme developed in T4ME/META-NET to provide a common metadata description for language resources

in many different European languages. Beside the general information about the language resource this metadata also covers information on IPR holders (the name of the holder as well as the addresses of the main contact person), as well as the licence issues and restrictions of its usage. The metadata also describes the NLP focused usage of the resources both in its actual and in its upcoming state (actual and foreseen usage). The metadata contains wider information of the resources by offering further readings and publications on the resources, as well as links of their main documentation. The metadata scheme of the resources also informs about data types as the media type of the resource or the language covered by the resource.

The description of resources in an uniform way gives a detailed view of the main language resources available on languages covered by the partners of the project - Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak. The focus was to gather all relevant information of the actually (or potentially) available resources.

#### WP2.3 Selection of resources of further interest (HASRIL, BME-TMIT, FFZG, IPIAN, ULodz, UBG, IPUP, IBL, LSIL)

Not all resources identified are in the particular focus of the CESAR project. In cooperation with the partner projects and META-NET the consortium defined the methodology and criteria to be used for a precise selection of resources and tools. Top-level criteria were availability, quality, standardization, quantity, usability, fitness, extensibility, perceived potential for reuse, recombination and repurposing. A particular place in the criteria list takes the estimation of the expected needs of different groups of end-users.

Based upon the agreed criteria and methodology the consortium is selecting the best possible mix of resources that will make the subject of further interest of different groups of end-users. Together with partner projects and META-NET the consortium ensures a balanced coverage of resources for different end-users and tasks, families of products and services, etc. The outline of the resources and tools with wide importance shows the possible gaps at the national and international levels and focuses the further efforts of the community. The selected resources of further interest will be presented in D2.5 – month 18th.

Submitted deliverables:

D2.1 Language Whitepaper, update M06

D2.2b Contact database of stakeholders, M18

D2.3c Report on resources (actually or potentially) available to the consortium, M18

D2.5 Report on resources of further interest, M18

#### WP3: Enhancing language resources (IPIAN)

In the second year of the project two subsequent batches of clean and reusable resources have been delivered (in July 2012 and January 2013) and made available through the open digital exchange META-SHARE provided by META-NET. The table below presents the statistics of these resources by partner, language and resource type (for convenience, the table gathers statistics for all 3 batches since numerous resources from batch 1 have been updated in batches 2 and/or 3):

	ILHAS	TMIT	FFZG	PIPAN	ULODZ	UBG	PUPIN	IBL	LSIL	Σ
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
<b>Total</b>	<b>31</b>	<b>25</b>	<b>26</b>	<b>59</b>	<b>17</b>	<b>19</b>	<b>2</b>	<b>36</b>	<b>36</b>	<b>251</b>

The delivered resources have been documented according to the metadata model made available by META-SHARE. Resource descriptions have been registered at CESAR partner nodes running the most recent version of the META-SHARE application.

META-SHARE node setup has been organized to maintain at least one META-SHARE node per language. All nodes are currently operational:

- 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/>

Following the META-SHARE tree-like structural organization of nodes, all resource descriptions from CESAR partner nodes are being harvested by the IPIPAN node. The IPIPAN node has also been promoted to the managing node status, which results in constant synchronization of resource descriptions with other managing nodes, thus propagating CESAR resource descriptions to the worldwide META-SHARE network.

To introduce additional level of security for critical resources, backup copies of selected resources have been transferred to IPIPAN disk matrix to be maintained throughout the sustainability period.

One of the auxiliary results of the workpackage was the design, implementation and delivery of the XSLT-based environment for generation of the human-readable descriptions of resources based on the META-SHARE metadata exported to XML (using standard META-SHARE functionality). This method was selected to avoid duplication of information already available in the metadata files and improve quality of the descriptions produced for META-SHARE.

The delivery of resources has been documented in the D3.1a, D3.2a and D3.3a deliverables, automatically generated using the above-mentioned framework. Actions performed on resources were summarized in deliverables D3.1b, D3.2b and D3.3b, containing additional explanations given by partners; below we provide a brief summary of work divided into three main categories: upgrade, extending and linking and aligning across languages.

WP3.1 Upgrading resources to agreed standards (HASRIL, BME-TMIT, FFZG, IIPAN, ULodz, UBG, IPUP, IBL, LSIL)

The upgrade task consisted of following actions:

- upgrade for interoperability (changing annotation format, type, tagset):
  - 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 lower-case characters, 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: 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): development 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,
  - 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, enhancement, conversion, standardization),
  - all resources: descriptions harmonized with META-SHARE metadata model.
  - Slovak National Corpus: individual document metadata have been converted to TEI P5
  - Manually Annotated Slovak Corpus: added bibliographic information (for human consumption)
  - Bulgarian wordnet: review and correction of synsets, literals, relations.
- harmonization of documentation (conversion to open formats, reformatting, linking)
  - 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.
  - Croatian National Corpus v3.0: new documentation on the corpus web site
  - Slovak National Corpus: comprehensive user manual, tutorial and documentation has been written
  - 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)
  - 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.
- other programming tasks (e.g. standardizing API calls):
  - 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: development 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:
  - 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 Newspapers 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.
- Multilingual 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.

### WP3.2 Extending and linking resources (HASRIL, BME-TMIT, FFZG, IPIPAN, ULodz, UBG, IPUP, IBL, LSIL)

The extension/linking task consisted of following actions:

- 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 million sentence pairs, respectively
  - Slovak Terminology database has been extended by several hundred 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 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.
- 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:
  - 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,
  - Hungarian Speech Emotion Database and Hungarian Telephone Client Speech Database: integration, merging wherever possible (in order to obtain an interlinked verbal+non-verbal corpus).

### WP3.3 Aligning resources across languages (HASRIL, BME-TMIT, FFZG, IPIAN, ULodz, UBG, IPUP, IBL, LSIL)

The cross-lingual alignment consisted of the following actions

- 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,
  - 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
  - alignment of Bulgarian, Croatian, Hungarian, Serbian and Slovak wordnets based on Princeton 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,
  - 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 Clausal-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.

The progress of the development in NooJ were made by IPUP in close collaboration with Max Silberztein, the author of original NooJ. The development 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 interoperable by making sure it will seamlessly work with major tools

Two platform independent versions of NooJ were developed during the project:

- MONO version of NooJ
- Java version of NooJ

Java version of NooJ is also an open source software, where NooJ GUI is separated from the engine thus supporting high interoperability.

Submitted deliverables:

- D3.2a – Second batch of language resources (part A): documentation of delivery, M18
- D3.2b – Second batch of language resources (part B): actions on resources, M18
- D3.3a – Third batch of language resources (part A): documentation of delivery, M24
- D3.3b – Third batch of language resources (part B): actions on resources, M24

#### WP4 Cross-national collaboration and Pilot service (BME-TMIT)

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

##### WP4.1 IPR and other legal issues (HASRIL, BME-TMIT, FFZG, IPIAN, ULodz, UBG, IPUP, IBL, LSIL)

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 second year of 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.

#### WP4.2 Harmonization of resource descriptions and related directory services (HASRIL, BME-TMIT, FFZG, IIPAN, ULodz, UBG, IPUP, IBL, LSIL)

Main aims were 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 during the project is as follows:

- Check and adapt metadata schemes, provide PSP wide feedback
- Work in close and continuous collaboration with other PSPs and META-NET, sharing of experiences, further emerging needs, problems, etc.
- Provide and/or harmonize and standardize metadata description for resources involved in the upload batches: include all mandatory as well as the most possible optional elements
- Upgrade all metadata from previous batches to latest versions for next batches

#### WP4.3 Population and pilot operation of the digital exchange platform (HASRIL, BME-TMIT, FFZG, IIPAN, ULodz, UBG, IPUP, IBL, LSIL)

Tasks until batch 2 involved pilot operation by setting up a physical exchange platform and several development nodes and keeping close touch with software developers. For batch 3, at least one node was created for each language, 8 Partners out of 9 run their own META-SHARE node, and are committed to run them beyond the end of the project.

Detailed tasks and their achievements during the two years of the project:

- 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 META-SHARE managing node and 7 META-SHARE network nodes
- 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

Submitted deliverables:

- D4.4 Second upload of language resources M18
- D4.5 Third upload of language resources M24

#### WP5 Outreach, awareness and sustainability (FFZG)

The main effort of this Work Package was to rise awareness about LT in respective countries and to prepare the project to become sustainable beyond the end of the EU-funded phase. 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.

The general description of these activities is presented here, but to avoid the repetition of the same information, the detailed lists of these activities such as participation at conferences/presentations/exhibitions/events, publications, media appearance, press releases, media coverage etc. are available in D5.2. All details on WP5 activities are listed in D5.1, D5.2, D5.3 and D5.4 and that they are omitted from these objectives to avoid of duplication of data.

#### WP5.1 Action plan for outreach, awareness and sustainability (HASRIL, FFZG, IPAN, IPUP, IBL, LSIL)

A detailed action plan for outreach, awareness and sustainability, that was developed at the beginning of the project, detailing awareness, community mobilisation and dissemination actions to be undertaken in each country covered by the project. It was updated in M12 to accommodate the harmonisation with the META-NET activities at large and to maximise the project's impact and ensuring its sustainability beyond the EU-supported phase. While in Y1 the action plan was focused on the first of the three main groups, i.e. on the research community in human language technology and other related domains, in Y2 the action plan and dissemination activities were oriented towards other two target audience groups, i.e. industry (both language industry and other business sectors) and society (.government and other public decision makers, as well as general public).

The analysis of users' needs, that were also performed in the WP2.1, served as valuable input that ensured that this action plan was tailored to the users' needs as much as possible.

The action plan was updated at M12 of the project to include further actions to be undertaken during the second year and after the end of the EU contract.

Since by the action plan the appropriate dissemination channels and events were defined at international and national level, in the second year we were following that plan using the series of potential contacts and channels for dissemination at national level and the lists of relevant players in LT field from the national level (research, business and policy stakeholders), collected within WP2 and provided by each partner in the CESAR consortium.

Upon harmonisation of visual identity elements with META-NET, we were exploiting these dissemination channels fully in scientific & non-scientific circles; printed & electronic publications, international events (science, technology, media, professional, industrial and political) as well as CESAR-specific events. Also, wherever and whenever needed, a joint META-NET dissemination strategy for incorporated appearance at major events (e.g. shared booths and production of dissemination materials at LREC2012) was discussed and developed with other members of META-NET Communication Working Group.

The public web site ([www.cesar-project.net](http://www.cesar-project.net)) was promoted to the main project dissemination channel and was fully treated as independent, yet harmonised and cross-linked with [www.meta-net.eu](http://www.meta-net.eu) web site. It was also visually linked to the META-NET alliance using common visual identity rules.

As planned, the most important means of enhancing awareness in business, society and government for CESAR project countries was a series of nationally organized high-level events („road shows“) that took place in each country. The original plan for organizing these events (starting with 2012-05 and ending in 2012-11) had to be rescheduled, but we were able to organize CESAR road-shows in each country. We found this format of events the most suitable for local governmental officials and industry leaders in getting acquainted with the META-NET Network of Excellence and the role of LRT in general, and national level stakeholders in particular. Expected impact was enhanced support for LRT at the national level by both, industry and government and this should lead to the support to the collected LRTs beyond the EC project funding, thus providing needed sustainability.

#### WP5.2 Mobilise the research community (HASRIL, BME-TMIT, FFZG, IPIAN, ULodz, UBG, IPUP, IBL, LSIL)

The research community at national level was reached by several dissemination channels (traditional or not so traditional) that were planned in D5.1 in order to attract the relevant players to participate in sharing resources and tools through META-SHARE platform. Also collaborative participation with national CESAR partner at different national and international conferences was favoured since this helped players at the national level and outside of CESAR consortium to present their work that might otherwise remain unseen. In this respect the CESAR project partners played the role of catalyst in transferring the information about different LT players from the national level to the European and global level. In this respect, including in Batch 1-3 LRTs from outside consortium was highly regarded and also represented a proof of good mobilisation activities at national level. Also, the role of CESAR consortium members was to convey information about the META-NET from European to the national level using defined dissemination channels.

Applicability of some type of seal of approval originated from international bodies/projects/initiatives/networks was also investigated as one of means of attracting LRT players from the national level. Such a seal (“Available on META-SHARE”) was suggested and agreed upon at the level of META-NET and the web page visual element has been designed. CESAR partners used this seal at their resource and tool web pages, but they will also promote its usage for certain nationally based or nationally funded language resources or tools that complies with predefined META-SHARE standard(s) for future inclusion of these LRTs in META-SHARE. This could also serve as one of means to attract other players at the national level and to mark certain level of compatibility and quality that resource or tool have.

#### WP5.3 Increase visibility within business and industry (HASRIL, BME-TMIT, FFZG, IPIAN, ULodz, UBG, IPUP, IBL, LSIL)

Industry at national level was reached by dissemination channels that were defined in WP5.1 in order to increase visibility of LT research. Each partner invited their local industrial stakeholders to participate in dissemination and demonstration actions/events at consortium (“road-shows”) and European level (e.g. META-FORUM in Bruxelles). The visibility within business and industry was combined with WP5.4, but was also highly dependent on the development of META-SHARE infrastructure that would be used for demonstration purposes.

By the end of the project, in D5.3b the final sustainability strategy and plans beyond the end of the project were presented.

#### WP5.4 Enhance awareness in society and government (HASRIL, BME-TMIT, FFZG, IPIAN, UBG, IPUP, IBL, LSIL; M08-M24):

Beside the defined dissemination channels, that were used for both, society and government awareness rising as usual, several awareness-rising activities were used at governmental/funding organisations/language councils levels (e.g. presentation of CESAR project at national ICT-PSP days, or at national science festivals or science days etc.).

Also, the international presentation of projects aims were strongly supported. The strategy we have put behind the “road-shows” was based on our experience that experts from abroad presenting the same topic to decision makers and founders at national level usually achieve much better results than domestic experts. This is why we developed a plan to organize a series of events in the form of a „road-show“ that would happen at the highest possible level in every CESAR country (preferably with participation at the ministerial or vice-ministerial level, which we successfully achieved in most of the cases). We expected that this kind of high-level presentation of META-NET and CESAR to all relevant stakeholders at national level (research, industry and policy) would yield the best possible results in rising awareness in governmental bodies and society in general. It was also planned and rightfully expected that officials from META-NET, DG Information Society & Media and later DG Connect, as well as other similar projects or institutions would provide support to this process by taking part at these events and giving presentations either in the capacity

of EC representatives or as foreign LRT experts giving examples of best practice in the neighbouring countries. This was the strategy that we followed in order to make stronger impact, raise awareness and visibility at governmental and funding agencies level and thus help local experts in gaining support for the LT field from national funding through clear and sound European support. These events were organized by local organizers, i.e. consortium member institutions from Bulgaria, Slovakia, Poland, Serbia, Croatia and Hungary that play the key role in LRT at their national level anyway. The logistics was centrally co-ordinated from the WP5 leading partner and executed at each national level.

The “road-show” events were organized at the following places and the final schedule was:

- Bulgarian road-show: Sofia, 2012-05-02
- Slovakian road-show: Bratislava, 2012-06-07 – 2012-06-08
- Polish road-show: Warsaw, 2012-09-27 – 2012-09-28
- Serbian road-show: Belgrade, 2012-10-29
- Croatian road-show: Zagreb, 2012-11-30
- Hungarian road-show: Budapest, 2013-01-18

and were supported by media coverage described in D5.4 deliverables, beside all other dissemination materials.

[WP5.5 EU-wide awareness, mobilisation and dissemination actions \(HASRIL, BME-TMIT, FFZG, IPIPAN, ULodz, UBG, IPUP, IBL, LSIL\)](#)

Project representatives actively participated in META-NET Communication Working Group, where they contributed to transnational events organized by the EC and/or META-NET e.g. for government and industry representatives adjoining the major scientific conferences or the annual event of the META-FORUM (2011 held in Budapest, 2012 held in Bruxelles).

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 participant at different events e.g. preparations of META-NET and CESAR dissemination materials for different conferences and other events.

The press releases were also carefully planned, translated and their issuing coordinated with the META-NET in order to achieve the maximum by a large number of national media appearances that finally influenced the European level i.e. interest by European commissioner(s) and EP members about the role of LT in multilingual Europe. This activities were organised around two important campaigns coinciding with the European Day of Languages (2012-09-26) and around 2013-01 when the META-NET Strategic Research Agenda was published. But at the same time, the CESAR “road-shows” happening in the autumn and winter 2012 benefited from this media appearances since at the national level the campaigns were also concentrated to maximise the visibility of the respective “road-shows”.

Submitted deliverables:

D5.4b Dissemination material, M18

D5.2.b Awareness, mobilisation and dissemination actions – annual report, M24

D5.3b Sustainability strategy and plans beyond the end of the project, M24