Central and South-East European Resources in META-SHARE

Tamás VÁRADI[1] *Marko TADIĆ*[2]
(1) RESERCH INSTITUTE FOR LINGUISTICS, MTA, Budapest, Hungary
(2) FACULTY OF HUMANITIES AND SOCIAL SCIENCES, ZAGREB UNIVERSITY, Zagreb, Croatia
varadi.tamas@nytud.mta.hu, marko.tadic@ffzg.hr

ABSTRACT

The purpose of this demo is to introduce the Language Resources, Tools and Services (LRTS) that are being prepared within the *Central and South-East European Resources* (CESAR) project. To the computational linguistic community the languages covered by CESAR (Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovakian) were so far considered under-resourced and their language resources and tools being insufficient or hard to obtain, with limited coverage and processing capabilities. The CESAR project, functioning as an integral part of META-NET initiative, aims to change this situation by coordinating all relevant national stakeholders, to enlarge and enhance the existing LRTSs, as well as connect them multi-lingually in various ways. The most important aim of the project is to make all LRTSs for respective languages accessible on-line through the common European LRTS distribution platform META-SHARE. This demo will present how this platform can be used in different scenarios and how researchers or industry partners can easily access CESAR Language Resources, Tools and Services.

KEYWORDS : language resources, language tools, language services, CESAR, META-NET, META-SHARE, Bulgarian, Croatian, Hungarian, Polish, Serbian, Slovakian

Proceedings of COLING 2012: Demonstration Papers, pages 431–438,
COLING 2012, Mumbai, December 2012.

1 Introduction

The purpose of the paper and the accompanying demonstration is to introduce the Language Resources, Tools and Services (LRTS) that are being prepared within the *Central and South-East European Resources* (CESAR) project. The CESAR project functions as an integral part of the larger, Europe-wide initiative META-NET[1], that tries to coordinate efforts for 30+ European languages in the field of LRTS. META-NET started as a network of excellence in 2010 and after a year it turned into a large META-NET initiative that encompasses and interlinks four EC-funded projects, resulting in a truly Europe-wide conglomerate of computational linguistic and NLP communities. CESAR is one of the projects involved. ***Section 2*** gives a brief description of the project within the META-NET context; ***section 3*** discusses its general aims and describes the META-SHARE [2] platform; ***section 4*** describes the CESAR results obtained so far and demonstrate their availability on-line; finally, the paper ends with a brief conclusion and suggestion for future development.

2 General CESAR description

CESAR stands for *Central and South-East European resources*, a CIP ICT-PSP-2010-4 *Theme 6: Multilingual Web* Pilot B type project, funded in 50:50% scheme by EC and national funding sources. The project started on 1st February 2011 and its duration is 24 months. The partners of the project are academic institutions coming from six Central and South-East European countries, namely, Bulgaria, Croatia, Hungary, Poland, Serbia and Slovakia, representing respective languages. Although some of these languages might be considered not to be completely under-resourced any more, still for most of them LRTSs have been developed mostly in a sporadic manner, in response to specific project needs, with relatively little regard to their long-term sustainability, IPR status, interoperability, reusability in different contexts as well as to their potential deployment in multilingual applications. In this respect, CESAR languages should be regarded as under-resourced languages and CESAR is aiming to change this situation.

3 Aims

High fragmentation and a lack of unified access to language resources are among key factors that hinder European innovation potential in language technology development and research. In that context the general aims of CESAR are coordinated with other projects in META-NET initiative.

3.1 Coordinating stakeholders

Even for languages with relatively well developed LRTSs, it is difficult or in many cases impossible to get access to resources that are scattered around different places, are not accessible online, reside within research institutions and companies and exist as “hidden language resources”, similar to the existence of the “hidden web”. CESAR wants to

[1] http://www.meta-net.eu.

[2] http://www.meta-share.eu

overcome this situation at the respective national levels by coordinating researchers, industrials and policy makers, and by putting them in their proper roles at the national LRTS landscape.

3.2 Actions with LRTSs in CESAR

In principle, the CESAR project doesn't produce new resources, but puts most of its efforts in their upgrading, extending and cross-lingual alignment.

The upgrade task mostly focuses on reaching META-SHARE compliance by upgrade for interoperability (changing annotation format, type, tagset), metadata-related work (creation, enhancement, conversion, standarization) and harmonization of documentation (conversion to open formats, reformatting, linking).

Existing resources are being 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. Probably the best example is merging of two pre-existing competitive Polish inflectional lexica (Morfeusz and Morfologik) with different coverage and encoding systems, into one large unified one (Polimorf Inflectional Dictionary).

Cross-lingual alignment of resources is the most demanding task and it will be applied only to a small number of resources close to the end of the project, mostly by producing collocational dictionaries and n-grams from national corpora using the common methodology.

3.3 META-SHARE

META-SHARE is a sustainable network of repositories of language data, tools and related web services documented with high-quality metadata, aggregated in central inventories allowing for uniform search and access to resources. Data and tools can be both open and with restricted access rights, free of charge or for-a-fee. META-SHARE targets existing but also new language data, tools and systems required for building and evaluating new technologies, products and services. In this respect, reuse, combination, repurposing and re-engineering of language data and tools play a crucial role.

META-SHARE is on its way to becoming an important component of a language technology marketplace for HLT researchers and developers, language professionals (translators, interpreters, localisation experts, etc.), as well as for industrial players that provide innovative HLT products and services to the markets.

META-SHARE users have a single sign-on account and are able to access everything within the repository. Each language resource has a permanent locator (PID). One of the key features of META-SHARE will be metadata harvesting, allowing for discovering and sharing resources across many repositories.

At the moment there are 1248 language resources, tools or services accessible through META-SHARE and they are distributed over 100+ languages, four main resource types (corpus, lexical/conceptual model, tool/service, language description) and four main media types (text, audio, image, video).

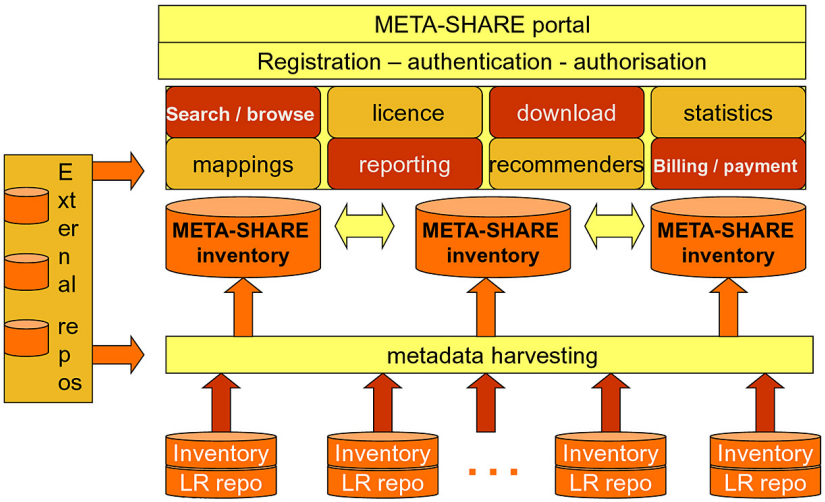

Figure 1: General structure of META-SHARE (Piperidis 2012)

Extensive usage of advanced metadata schemata for description enables automatic harvesting and discovery of resources within the network of repositories (Gavrilidou et al. 2012).

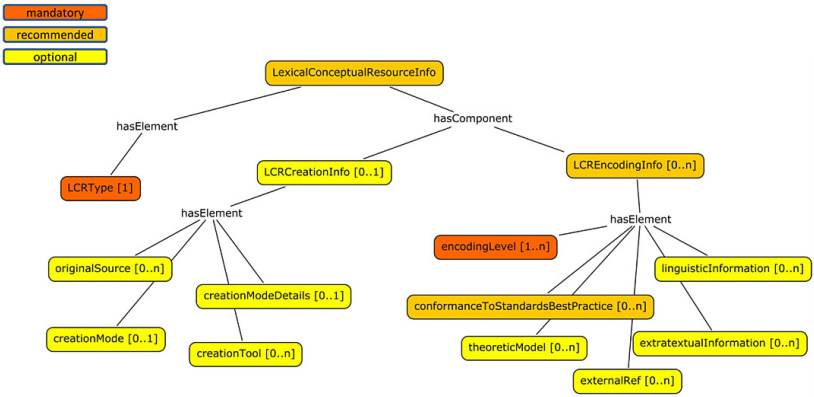

Figure 2: Metadata schema for Lexical/Conceptual resource (Monachini, 2011)

4 CESAR LRTSs in META-SHARE

The description of CESAR resources were prepared in compliance with the META-SHARE component-based metadata model. The taxonomy of LRTSs includes two-level hierarchy, with general main resource type and type-dependent subclassification. The metadata were prepared by CESAR partners with the intention of providing detailed informaton on each LRTS. It is planned that CESAR LRTSs will be submitted to META-SHARE in three separate batches. So far for the first two batches all relevant metadata have been uploaded in November 2011 and July 2012, with the third planned for January 2013.

Currently, after two batches in CESAR META-SHARE node 127 CESAR LRTSs are accessible, out of which there are 68 corpora, 16 dictionaries, 3 lexicons, 4 wordnets, 8 speech databases and 28 tools, but the number is changing with each new LRTS made accessible through this platform. To illustrate the size of LR in cumulative numbers over six CESAR languages, what is accessible now encompasses 1.7 billion tokens in monolingual corpora, 41.8 million tokens in parallel corpora, and 1.6 million lexical entries/records.

Within the META-SHARE platform a specialised editor for entering metadata about individual LRTS was developed. It enables finetuning the metadata about each resource.

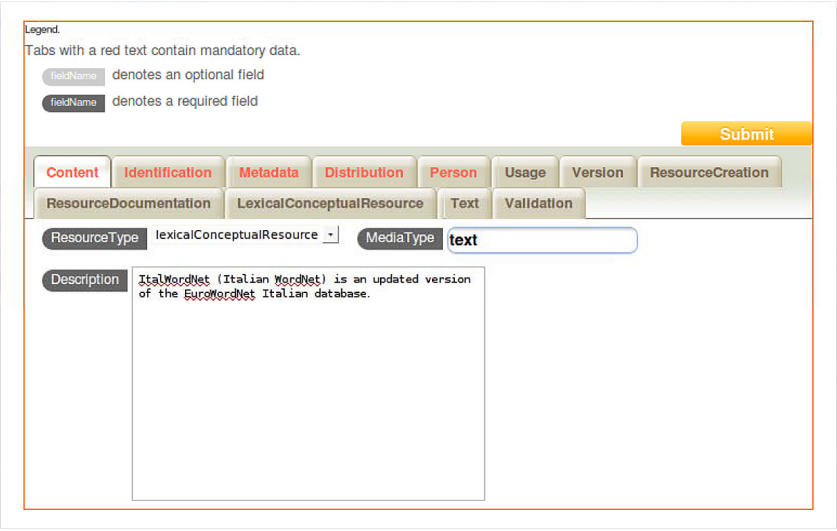

Figure 3: Entering metadata about new resource using META-SHARE editor

However, to speed up the development time, CESAR metadata descriptions were prepared in XML format off-line in accordance with the predefined schema and uploaded into the CESAR META-SHARE node, currently operated by IPIPAN, Warsaw for all CESAR partners. Referenced resources are stored by their respective owners.

Once the metadata about LRTS is stored, META-SHARE enables the user to use a search function, so that users can search and browse through the network of repositories in the most flexible way possible, using the panel with different filtering criteria available on the left hand side of the META-SHARE website window. An overall search engine is also available on the top.

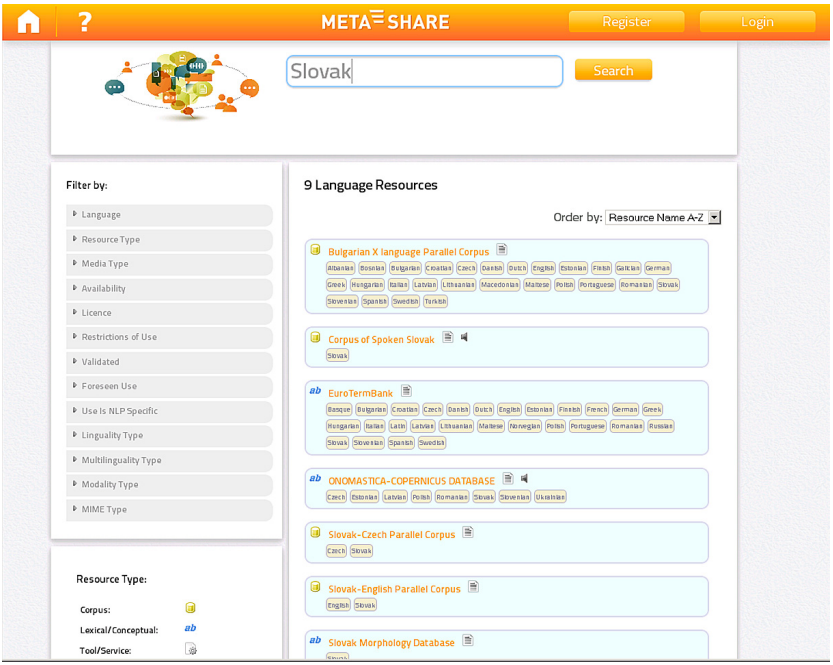

Figure 4: Search results of a query in META-SHARE

It is our intention to demonstrate to the research community in computational linguistics and NLP the features and possibilities of META-SHARE platform, particularly stressing the visibility and accessibility of LRTSs for six Central and South-East European under-resourced languages.

Conclusions and future development

With this demo paper we intend to demonstrate to the computational linguistic community how simple it has become to find in the META-SHARE platform language resources, tools and services for languages covered by CESAR that are usually considered under-resourced or at least not easy to access. The future steps in the CESAR project will include the uploading of the final batch or LRTSs and particularly publishing of NooJ, the open source NLP development environment, newly ported into JAVA under the aegis of the CESAR project.

Acknowledgment

This paper presents work done in the framework of the project CESAR, funded by DG INFSO of the European Commission through the ICT-PSP Program, Grant agreement no.: 271022.

References

Federmann, C., Giannopoulou, I., Girardi, C., Hamon, O., Mavroeidis, D., Minutoli, S., Schröder, M. *META-SHARE v2: An Open Network of Repositories for Language Resources including Data and Tools*. LREC2012, (pp. 3300-3303).

Garabík, R., Koeva, S., Krstev, C., Ogrodniczuk, M., Przepiórkowski, A., Stanojević, M., Tadić, M., Váradi, T., Vicsi, K. Vitas, D. & Vraneš, S. (2011) CESAR resources in META-SHARE repository. LTC2011 (pp. 583).

Gavrilidou, M., Labropoulou, P., Desipri, E., Giannopoulou, I., Hamon, O. & Arranz, V. (2012) *The META-SHARE Metadata Schema: Principles, Features, Implementation and Conversion from other Schemas*. Workshop "Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR". LREC2012 (pp. 5-12).

Lyse, G. I., Desipri, E., Gavrilidou, M., Labropoulou, P., Piperidis, S. (2012) *META-SHARE overview*. Workshop on the Interoperability of Metadata, Oslo, 2012-06-05.

Monachini, M. (2011) *Metadata for Lexical and Conceptual Resources*. Athens Workshop IPR-Metadata, Athens, 2011-10-10/11.

META-SHARE documentation & user manual (2012) [http://www.meta-net.eu/meta-share/].

META-SHARE knowledge base (2012) [http://metashare.ilsp.gr/portal/knowledgebase].

Piperidis, S. (2012) The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions. LREC2012, (pp. 36-42)