CESAR resources in META-SHARE repository

Radovan Garabík[1], Svetla Koeva[2], Cvetana Krstev[3], Maciej Ogrodniczuk[4], Piotr Pęzik[5], Adam Przepiórkowski[4], Mladen Stanojević[6], Marko Tadić[7], Tamás Váradi[8], Klára Vicsi[9], Duško Vitas[3], Sanja Vraneš[6]

[1]Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences, [2]Institute for Bulgarian Language, Bulgarian Academy of Sciences, [3]University of Belgrade, [4]Institute of Computer Science, Polish Academy of Sciences, [5]University of Łódź, [6]Institut Mihajlo Pupin, [7]University of Zagreb, [8]Research Institute for Linguistics, Hungarian Academy of Sciences, [9]Budapest University of Technology and Economics

Abstract

The aim of this demo is to present multilingual resources made available in the new META-SHARE open infrastructure by partners of the CESAR consortium (CEntral and South-east europeAn Resources, a European CIP ICT-PSP project, Grant Agreement 271022, `http://www.cesar-project.net`) in November 2011, within the first batch of resources to be delivered in 2011-2013.

Keywords: META-SHARE, META-NET, CESAR, language resources and tools, LRT, natural language processing, NLP

1 Introduction

The CESAR project, part of the META-NET Network of Excellence (Multilingual Europe Technology Alliance, see `http://www.meta-net.eu`), is targeted to deliver clean and reusable language resources through the open digital exchange provided by META-NET (called META-SHARE) or other suitable channels. Languages in scope of the project and represented by the consortium partners are Bulgarian, Croatian, Hungarian, Polish, Serbian and Slovak.

2 Enhancing language resources

The preparation of resources for the first batch followed the first two of the three main activities:

1. upgrading resources to agreed standards,
2. extending and linking resources,
3. aligning resources across languages.

The upgrade task mostly focused 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 were extended/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.

Cross-lingual alignment of resources, as the most demanding task, will be applied only to a small number of resources in the next batches, planned for July 2012 and January 2013.

3 The metadata model

The descriptions of CESAR resources were prepared in compliance with the META-SHARE component-based metadata model. The taxonomy of language resources includes two-level hierarchy, with general "main type" classification (corpus, lexical/conceptual resource, language description or technology/tool) and type-dependent subclassification.

The metadata descriptions were prepared by CESAR partners with intention of providing detailed information on each resource (conforming to META-SHARE *maximal schema*). The descriptions were then used for automated, XSLT-based documentation generation.

4 META-SHARE repository

The META-SHARE is an open distributed facility for sharing and exchange of resources developed by META-NET. META-SHARE servers offering access to resources are intended to run as network nodes, synchronizing metadata descriptions and maintaining access permissions.

Currently the basic release (version 1) of the META-SHARE application offers metadata maintenance (in a Web-based editor) as well as import and export; the open source release with extended functionality is planned for January 2012 and community release for July 2012.

5 First batch resources

The following 33 resources for 6 languages have been made available in the first batch:

- 20 corpora (19 written, 1 multimodal),
- 2 dictionaries,
- 4 wordnets,
- 1 lexicons,
- 6 speech databases.

All resources underwent conversion for standardization, careful metadata description, documentation update and licensing clarification. Their metadata descriptions were prepared in XML format and uploaded into the CESAR META-SHARE node (based in Warsaw), common for all CESAR partners. Referenced resources are stored by their respective owners.