Language resources, tools and services in the CESAR project and META-* framework

Marko Tadić

University of Zagreb
Faculty of Humanities and Social Sciences
Zagreb, Croatia
marko.tadic@ffzg.hr

Informacijska družba 2012 – Jezikovne tehnologije
Ljubljana, Slovenia, 2012-10-08
Outline

- META-NET
  - META-SHARE platform and Charter
  - Language Whitepapers
- CESAR project in general
  - geo-linguistic spread, partners in the consortium, general aims
  - resources, tools and services
  - NooJ@CESAR
- conclusions
META-NET, META-SHARE, META
META-NET

- Network of Excellence consisting of
  - 60 research centres
  - 34 countries
- dedicated to build the technological foundations of a multilingual European information society
- all EU member states and several non-member states covered
META

- META-NET is the core of META (Multilingual Europe Technology Alliance)
  - 642 members, 55 countries
  - including W3C, Google, Microsoft, GALA, research centres, LT companies etc.
- META brings together
  - researchers, commercial technology providers, private and corporate
    language technology users, language professionals, other information
    society stakeholders
- META prepares an ambitious, joint international effort
  - Language Technology as one of means for realising Europe united as one
    single digital market and information space
META-NET and ICT-PSP projects

- META-NET (www.meta-net.eu)
  - T4ME: Network of excellence (FP7), IJS one of project partners
- supported by three additional projects (ICT-PSP)
  - META-NORD (www.meta-nord.eu)
  - META-NET4U (www.metanet4u.eu)
  - CESAR (www.cesar-project.net)
- duration of all ICT-PSP projects from 2011-02-01 to 2013-01-31
- META-NET = coordinated pan-European action
  - covering 30 European languages
- joint effort to harmonise Language Resources, Language Tools and Language Services (LRTS) for all these languages in order to achieve
  - sustainability, interoperability, reusability
META-SHARE

- open exchange infrastructure for language resources and tools
- language resources and tools are documented, uploaded, stored in repositories, catalogued, can be downloaded, shared, discussed
- improve their visibility, documentation, identification, availability, preservation, interoperability
- long-term goal: boost research, technology and innovation through wide availability, pooling, openness and sharing of resources
- repositories store and maintain inventories of resources and tools
- metadata inventories are exported and harvested in the network
- currently 13 repositories up and running; >1,300 LRs available
- www.meta-share.eu
META-SHARE architecture

META-SHARE portal
Registration - Authentication - Authorisation

- Search / Browse
- License
- Download
- User support
- Mappings
- Reporting/Statistics
- Recommendations
- Billing / Payment

META-SHARE inventory

Metadata harvesting

External repos

Inventory LR repo

Inventory LR repo

Inventory LR repo

Inventory LR repo
1,359 language resources at your disposal...

Type in your keywords, please...

What is it? - About the project

META-SHARE, the open language resource exchange facility, is devoted to the sustainable sharing and dissemination of language resources (LRs) and aims at increasing access to such resources on a global scale.

META-SHARE is an open, integrated, secure and interoperable sharing and exchange facility for LRs (datasets and tools) for the Human Language Technologies domain and other applicative domains where language plays a critical role.

META-SHARE is implemented in the framework of the META-NET Network of Excellence. It is designed as a network of distributed repositories of LRs, including language data and basic language processing tools (e.g., morphological analysers, PoS taggers, speech recognisers, etc.).
11 Language Resources

Order by: Resource Name A-Z

- ACCURAT balanced test corpus for under resourced languages
  - Language: {croatian, english, estonian, german, greek, latvian, lithuanian, romanian, slovenian}
  - Language: {croatian, english, estonian, german, greek, latvian, lithuanian, romanian, slovenian}
  - Language: {croatian, english, estonian, german, greek, latvian, lithuanian, romanian, slovenian}
  - Language: {croatian, english, estonian, german, greek, latvian, lithuanian, romanian, slovenian}
  - Language: {croatian, english, estonian, german, greek, latvian, lithuanian, romanian, slovenian}
  - Language: {croatian, english, estonian, german, greek, latvian, lithuanian, romanian, slovenian}
  - Language: {croatian, english, estonian, german, greek, latvian, lithuanian, romanian, slovenian}
  - Language: {croatian, english, estonian, german, greek, latvian, lithuanian, romanian, slovenian}
  - Language: {croatian, english, estonian, german, greek, latvian, lithuanian, romanian, slovenian}
  - Language: {croatian, english, estonian, german, greek, latvian, lithuanian, romanian, slovenian}

- ACCURAT Bilingual Comparable Corpora for under-resourced languages
  - Language: {croatian, english, estonian, german, greek, latvian, lithuanian, romanian, slovenian}
  - Language: {croatian, english, estonian, german, greek, latvian, lithuanian, romanian, slovenian}
  - Language: {croatian, english, estonian, german, greek, latvian, lithuanian, romanian, slovenian}
  - Language: {croatian, english, estonian, german, greek, latvian, lithuanian, romanian, slovenian}
  - Language: {croatian, english, estonian, german, greek, latvian, lithuanian, romanian, slovenian}
  - Language: {croatian, english, estonian, german, greek, latvian, lithuanian, romanian, slovenian}
  - Language: {croatian, english, estonian, german, greek, latvian, lithuanian, romanian, slovenian}
  - Language: {croatian, english, estonian, german, greek, latvian, lithuanian, romanian, slovenian}
  - Language: {croatian, english, estonian, german, greek, latvian, lithuanian, romanian, slovenian}
  - Language: {croatian, english, estonian, german, greek, latvian, lithuanian, romanian, slovenian}

- ACCURAT corpus of comparable sentences
  - Language: {croatian, english, estonian, german, greek, latvian, lithuanian, romanian, slovenian}

- ACCURAT corpus of Wikipedia texts
  - Language: {croatian, english, estonian, german, greek, latvian, lithuanian, romanian, slovenian}

- EuroTermBank
  - Language: {croatian, english, estonian, german, greek, latvian, lithuanian, romanian, slovenian}

- INTERA Corpus
  - Language: {croatian, english, estonian, german, greek, latvian, lithuanian, romanian, slovenian}

- LC-STAR English-Slovenian Bilingual Aligned Phrasal lexicon
  - Language: {croatian, english, estonian, german, greek, latvian, lithuanian, romanian, slovenian}
Slovenian BNSI Broadcast News Speech Corpus

This speech database consists of TV news shows (both evening news, "TV Dnevnik" and late night news, "Odmevi"), from the archive of a Slovenian national broadcaster RTV Slovenia. The recordings took place between June 1999 and May 2003. The database comprises a total of 36 hours of recordings (… Read More)

Distribution
- Availability: Available - Restricted Use
- License: ELRA_VAR

Membership
- Member: True

Institution: ELRA
- License: ELRA_END_USER

License
- Member: True
- Membership institution: ELRA

Metadata
- Metadata creation date: 2005-05-12

Corpus audio
- Media: audio

Linguality
- Language type: Monolingual
- Language: Slovenian

Audio size
- Size: no size available
- Size unit: Other
META-SHARE: current contents

Resource types:
- corpus: 586
- lexicalConceptualResource: 650
- toolService: 294
- languageDescription: 20

Distribution per Language:
- English: 293
- Spanish: 503
- French: 391
- German: 40
- Italian: 36
- Chinese: 21
- Dutch: 28
- Portuguese: 25
- Greek: 22
- Swedish: 76
- Catalani: 41
- Arabic: 99
- Japanese: 41
- Romanian: 204
- Danish: 202
- Other: 30
- Polish: 99

META-SHARE Charter

- a demand for free usage of digital sources for research purposes
- http://www.cesar-project.net/metashare/METASHARE_Charter.pdf
- “any grant of access to LRs should include at least the right to both humans and machines to read the relevant content”
- “allowing transformative uses, dissemination and distribution of such resources is strongly encouraged”
- “the restrictions to use or re-use LRs should be the minimum possible so that
  - LRs are continuously enriched
  - language technologies and related services are fully deployed and further developed”

http://www.cesar-project.net
META-NET: Language Whitepapers

- major achievement so far: META-NET Language White Paper Series
  - 30 languages described in an uniform way
  - providing information on language
    - structure
    - history and standardization process
    - presence in (multi)media and Internet
    - status of language technologies
    - availability of language resources
    - development of language tools
  - addressing journalists and politicians
- this snapshot gives a comparative insight
  where are gaps for particular languages
  - directions for LRT to be developed
30 Languages Covered

- Basque
- Bulgarian*
- Catalan
- Croatian
- Czech*
- Danish*
- Dutch*
- English*
- Estonian*
- Finnish*
- French*
- Galician
- German*
- Greek*
- Hungarian*
- Icelandic
- Irish*
- Italian*
- Latvian*
- Lithuanian*
- Maltese*
- Norwegian
- Polish*
- Portuguese*
- Romanian*
- Serbian
- Slovak*
- Slovene*
- Spanish*
- Swedish*

* Official EU language
LWP: few figures

- >200 national experts contributed as authors or co-authors (ca. 5 per language on average)
- >50 additional experts have contributed data and information
- >8,000 copies were printed and distributed by META-NET
- Europe-wide campaign started
  - more than 200 media appearances in media around Europe and globe

http://www.cesar-project.net
Europe’s Languages and LT

- English
- Dutch
- French
- German
- Italian
- Spanish
- Catalan
- Czech
- Finnish
- Hungarian
- Polish
- Portuguese
- Swedish
- Basque
- Bulgarian
- Danish
- Galician
- Greek
- Norwegian
- Romanian
- Slovak
- Slovene
- Croatian
- Estonian
- Icelandic
- Irish
- Latvian
- Lithuanian
- Maltese
- Serbian

good support through Language Technology

weak or no support
CESAR project
Geo-linguistic position

- CESAR stands for CEntral and Southeast EuropeAn Resources
- CESAR operates as a part of META-NET Network of Excellence
- one of three supporting ICT-PSP projects defined with their geo-linguistic spread
  - Central and Southeast Europe
  - three inner seas: Baltic, Adriatic, Black Sea
- CESAR covers languages:
  - Polish EU, 38M (40-48M)
  - Slovak EU, 5.4M (7M)
  - Hungarian EU, 10M (16M)
  - Croatian EU in 2013, 4.4M (5.5M)
  - Serbian candidate soon, 7.3M (9M)
  - Bulgarian EU, 7.5M (9M)
- all languages Slavic, except Hungarian
CESAR Consortium

- **Bulgaria**
  - Bulgarian Academy, Institute for Bulgarian Language L. Andreychev

- **Croatia**
  - University of Zagreb, Faculty of Humanities and Social Sciences

- **Hungary**
  - Hungarian Academy, Research Institute for Linguistics
  - Budapest University of Technology and Economics

- **Poland**
  - Polish Academy of Sciences, Institute of Computer Science
  - University of Łódź

- **Serbia**
  - University of Belgrade, Faculty of Mathematics
  - Institute Mihajlo Pupin

- **Slovakia**
  - Slovak Academy of Sciences, L. Štúr Institute of Linguistics
Latest news

CESAR @ FASSBL in Dubrovnik/Croatia, September 19–21th, 2012

At the 8th FASSBL conference CESAR project was presented with several papers, namely:

1. Nikola Ljubašić: Comparing Traditional and Web Corpora of Croatian
2. Danijela Meherer, Željko Agić, Marko Tadić: Automatic Enrichment of Croatian Morphological Lexicon Using Large Corpora and Web Search
3. Krešimir Božet, Matea Srebačić, Vanja Štefanec: Derivational Patterns of Croatian Verbs
4. Čvetana Kršev, Đuško Vitas, Milos Utić: Derivational Patterns in E-Dictionaries of Serbian

This bi-annual conference held from 19th to 21st Sept. 2012 in Dubrovnik, Croatia, encompasses contributions covering South-Slavic and Balkan languages treated with formal linguistic and computational approaches. Two CESAR partners were present: FFZG and UBD.

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

- Language resources & tools (LRT) in CESAR countries were developed:
  - mostly in a sporadic manner
  - according to specific project needs
  - with little or no regard to:
    - long-term sustainability
    - IPR status
    - interoperability
    - reusability in different contexts (e.g. in multilingual applications)

- CESAR project aims to address this issues by:
  - enhancing and upgrading
  - standardising
  - cross-linking

  a wide variety of language resources and tools

- Making these LRT available through META-SHARE platform
General aims 2

- Resources will include mono and multilingual
  - Corpora
  - Dictionaries and wordnets
  - Relevant LT processing tools
    - Tokenisers
    - Lemmatisers
    - Taggers
    - Chunkers and parsers
  - Speech databases
- These efforts are directed to ensure sustainability of LT through
  - Mobilising the national LT and other research communities
  - Raising awareness of the role of language resources amongst
    - R&D policy makers
    - Media
    - General public

http://www.cesar-project.net
NooJ in CESAR

- **NooJ development environment**
  - http://www.nooj4nlp.net
  - it will play a significant role in rising the popularity of LT
  - corpus manager easy to implement and use
  - based on the widened concept of local grammars (also CF & CS grammars)
  - support for five CESAR languages already (lexicons, grammars etc.)
  - selected in CESAR as a showcase how multilingual and multilevel processing tools can be developed and applied to all languages

- CESAR will make NooJ **open source software** available for all platforms
  - Windows
  - Linux, Unix
  - Mac

  - NooJ tutorial
CESAR Batch 1 (2011-11)

- **Bulgarian**
  - Bulgarian National Corpus, Bulgarian National Corpus Collocation service, Bulgarian Part-of-Speech Corpus, Bulgarian Sense-annotated Corpus, Bulgarian X language Parallel Corpus, Bulgarian wordnet, Wordnet web service, Bulgarian Spell Checker for Windows, Bulgarian Spell Checker Web Service

- **Croatian**
  - Croatian National Corpus (HNK), Croatian Morphological Lexicon (HML), Croatian-English Parallel Corpus, Croatian Lemmatisation server, Croatian Vallency Dictionary (CROVALLEX)

- **Hungarian**
  - Szeged Corpus 2.0, Szeged Treebank 2.0, Szeged Named Entity Recognition Corpus, Hungarian WordNet, Hungarian Webcorpus, Hunglish Parallel Corpus Version 2.0, morphdb.hu, hunmorph tool, hunalign tool, huntoken tool, Mindentudás Speech Corpus, Word Level Speech Database, Hungarian BABEL, Hungarian Broadcast News Database, Sound Gesture Database, Hungarian Speech Emotion Database

http://www.cesar-project.net
CESAR Batch 1 (2011-11)

- **Polish**
  - PoliMorf Inflectional Dictionary, Polish WordNet, Polish Named Entity Recognition Tool, 1 million subcorpus of National Corpus of Polish, Polish Named Entity Resources, LUNAPL and LUNA-WOZPL corpora, PELCRA Polish-English parallel corpora, PELCRA Polish spoken corpus

- **Serbian**

- **Slovak**
  - Slovak National Corpus, Corpus of Spoken Slovak, Slovak Morphology Database, Slovak-Czech Parallel Corpus, Slovak-English Parallel Corpus

http://www.cesar-project.net
CESAR Batch 2 (2012-07)

- **Bulgarian**
  - Bulgarian-X Language Parallel Corpus Collocation service, Lists of Bulgarian Multiword Expressions, Bulgarian Frequency Dictionary, Hydra - tool for developing wordnets, Chooser – annotation tool, Bulgarian Sentence splitter and Tokenizer, Web based infrastructure for Bulgarian data processing

- **Croatian**
  - Croatian Web Corpus, Slovene Web Corpus, Croatian-English Parallel Web Corpus, South-East European Parallel Corpus, Croatian Dependency Treebank, Web Content Extractor, Collocation and Term Extractor

- **Serbian**
  - English-Serbian Aligned Corpus, Serbian NooJ module, Serbian Morphological Dictionary (Multext-East), Morphosyntactically tagged Serbian version of Jules Verne's novel "Around the world in 80 days", Bibliša: Aligned Collection Search Tool, Corpus of Contemporary Serbian Newspapers and Magazines

- **Slovak**
Hungarian

- SzegedParallel, SzegedParallel FX, Szeged Treebank FX, Hungarian WSD Corpus, Szeged Criminal NE Corpus, Hungarian Verb Phrase Constructions, Hungarian NER Corpus based on Wikipedia, Hungarian Opinion-Tagged Sentence Bank, Hungarian Language Processing Tools in NooJ, hunner tool, hunpars tool, hunpos tool, Hungarian MTBA, Hungarian MRBA, Hungarian Telephone Speech Call Center Database, Hungarian BABEL phonetic and prosodic segmentation and syntactic analysis, Di-phone database for text-to-speech conversion in Hungarian, Hungarian Read Speech Precisely Labelled Parallel Speech Corpus Collection, Read Speech database in Hungarian, Hungarian Book (Egri csillagok/Eclipse of the Crescent Moon by Géza Gárdonyi) – a Read Speech and Aligned Text Selection Database, Hungarian Poem (János vitéz/John the Valiant by Sándor Petőfi) – a Read Speech and, Aligned Text Selection Database, Hungarian Parliamentary Speech and Aligned Text Selection Database,
CESAR Batch 2 (2012-07)

- **Polish**
  - Polish Sejm Corpus, Morphosyntactic tagset converter for positional tagsets, Spejd, N-grams from balanced National Corpus of Polish, Distributable subcorpus of National Corpus of Polish, Morfeusz PoliMorf, Morfologik Inflectional Dictionary, Morfologik-stemming, Grammatical Lexicon of Polish Phraseology, Grammatical Lexicon of Polish Economical Phraseology, Grammatical Lexicon of Warsaw Urban Proper Names, Multilingual lexicon of toponyms, Polish Valence Dictionary, Summarizer, Corpus of the Polish language of the 1960s, Shallow Grammar for the National Corpus of Polish, PANTERA, PolNet, ECL Dictionaries, PELCRA EN Lemmatizer, PELCRA Language Detector, PELCRA Polish-English parallel corpus of literary works (CC-BY), PELCRA multilingual parallel corpora (CC-BY), OSW Polish-English parallel corpus (CC-BY-NC), PELCRA time-aligned spoken corpus of Polish (CC-BY-NC), PELCRA WebLign crawler, PELCRA Word Aligned Corpora
Batch 1&2: quick statistics

- monolingual corpora
  - 1,702,656,806 token
- parallel corpora
  - 41,810,000 token
- lexicon entries/records
  - 1,640,579
- divided between
  - 32 corpora
  - 12 lexical resources
  - 20 tools/services

http://www.cesar-project.net
## CESAR Batch 1-3: total resources

<table>
<thead>
<tr>
<th></th>
<th>HU</th>
<th>CR</th>
<th>PL</th>
<th>RS</th>
<th>BG</th>
<th>SK</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HASRIL</strong></td>
<td>18</td>
<td>11</td>
<td>8</td>
<td>12</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td><strong>BME-TMIT</strong></td>
<td>13</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>3</td>
<td>90</td>
</tr>
<tr>
<td><strong>Corpus</strong></td>
<td>8</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>40</td>
</tr>
<tr>
<td><strong>Lexical resource</strong></td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>6</td>
<td>19</td>
<td>60</td>
</tr>
<tr>
<td><strong>Tool, service</strong></td>
<td>6</td>
<td>6</td>
<td>3</td>
<td>18</td>
<td>31</td>
<td>22</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>54</td>
<td>21</td>
<td>41</td>
<td>18</td>
<td>31</td>
<td>22</td>
</tr>
</tbody>
</table>

[http://www.cesar-project.net](http://www.cesar-project.net)
## Distribution of license types

<table>
<thead>
<tr>
<th>License</th>
<th>Quantity</th>
<th>License</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGPL</td>
<td>1</td>
<td>ELRA_END_USER</td>
<td>1</td>
</tr>
<tr>
<td>Apache License v2.0</td>
<td>2</td>
<td>GPL</td>
<td>12</td>
</tr>
<tr>
<td>BSD-style</td>
<td></td>
<td>LGPL</td>
<td>6</td>
</tr>
<tr>
<td>CC_BY</td>
<td>14</td>
<td>LGPL v3</td>
<td>2</td>
</tr>
<tr>
<td>CC_BY-NC</td>
<td>11</td>
<td>MSCommons_BY-NC-SA</td>
<td>1</td>
</tr>
<tr>
<td>CC_BY-NC-ND</td>
<td>1</td>
<td>MSCommons_COM-NR-ND-FF</td>
<td>3</td>
</tr>
<tr>
<td>CC_BY-NC-SA</td>
<td>2</td>
<td>MSCommons_NoCOM-NC-NR</td>
<td>10</td>
</tr>
<tr>
<td>CC_BY-NC-SA_3.0</td>
<td>2</td>
<td>other</td>
<td>24</td>
</tr>
<tr>
<td>CC_BY-SA</td>
<td>12</td>
<td>proprietary</td>
<td>6</td>
</tr>
<tr>
<td>CLARIN_ACA-NC</td>
<td>3</td>
<td>underNegotiation</td>
<td>2</td>
</tr>
<tr>
<td>CLARIN_RES</td>
<td>4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusions
Conclusions

- META-NET framework
  - incl. META-SHARE and Language Whitepapers series

- CESAR project
  - incl. NooJ as an open-source development environment

- Language resources, tools & services
  - incl. batches 1 & 2
  - all available through META-SHARE
  - under different licenses
Thank you for your attention.

http://www.cesar-project.net

office@meta-net.eu

http://www.facebook.com/META.Alliance