The CESAR Project: Enabling LRT for 70M+ Speakers

Marko Tadić

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

META-FORUM 2011
Budapest, Hungary, 2011-06-28

Co-funded by the 7th Framework Programme of the European Commission through the contract T4ME, grant agreement no.: 249119.

Co-funded by the ICT PSP Programme of the European Commission through the contract CESAR, grant agreement no.: 271022.

Outline

- CESAR project in general
  - geo-linguistic spread
  - partners in the consortium
  - general aims
- brief overview of situation in three countries
  - Croatia
  - Serbia
  - Slovakia
- conclusions

CESAR project

Geo-linguistic position

- CESAR stands for **CE**ntral and **S**outheast Europe**A**n **R**esources
- CESAR operates as a part of META-NET NoE
- 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, Ľ. Štúr Institute of Linguistics

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 interoperable mono and multilingual
  - speech databases
  - corpora
  - dictionaries and wordnets
  - relevant LT processing tools
    - tokenisers
    - lemmatisers
    - taggers
    - chunkers and parsers
- effort will be made to ensure sustainability through
  - mobilising the national LT communities
  - raising awareness of the role of language resources amongst
    - R&D policy makers
    - media
    - general public

Croatia

Croatia: Research in LT

- **University of Zagreb, Faculty of Humanities & Social Sciences**
  - long tradition: the first Croatian computational corpus, Bujas *Osman*, 1967
    - Croatian Frequency Dictionary (1999) on the basis of 1M-corpus of Croatian literary language (1976-1996)
  - today
    - Croatian National Corpus, since 1998, http://hnk.ffzg.hr
    - Croatian-English Parallel Corpus, since 2000
    - Croatian WordNet, since 2007, http://rmjt.ffzg.hr/p3.html
    - Croatian Dependency Treebank, since 2007, http://hobs.ffzg.hr
    - Croatian Morphological Lexicon/Lemmatisation Server, 2003, http://hml.ffzg.hr
    - CroTag, hybrid MSD-tagger (MulText East compliant), since 2006
    - Croatian NERC system, since 2005
    - Croatian module for NooJ, 2009
  - projects
    - national: Computational Linguistic Models & LT for Croatian, http://rmjt.ffzg.hr
    - bilateral: CADIAL – joint Flemish-Croatian project, http://www.cadial.org
    - EU: TELRI I & II, CLARIN, ACCURAT, LetsMT!, XLIKE

Croatia: Research in LT 2

- **Institute of Croatian Language and Linguistics**
  - Croatian Language Repository, since 2005, http://riznica.ihjj.hr
  - terminological databases, http://struna.ihjj.hr
  - digital dictionaries of Croatian dialects (incl. geo-mapping)
- **University of Zagreb, Faculty of Electrical Engineering and Computing**
  - Hascheck, on-line spelling checker, since 1994, http://hacheck.tel.fer.hr
  - Knowledge Technologies Laboratory, http://ktlab.fer.hr
    - information retrieval, information extraction
    - knowledge technologies, visualisation
    - tools: CorAl (corpus aligner), TermeX (terminology extraction)
  - projects
    - national: Knowledge discovery in textual data, http://rmjt.ffzg.hr/p5.html
    - AIDE – Automatic Indexing of Documents with Eurovoc, http://hidra.srce.hr:8080/eCadis/eCadis.jsp
    - bilateral: CADIAL – joint Flemish-Croatian project, http://www.cadial.org

Croatia: Research and beyond

- **University of Rijeka**
  - speech processing unit
  - Croatian spoken corpus
- association and portal
  - Croatian Language Technologies Society, since 2004, http://www.hdjt.hr
  - portal Language Technologies for Croatian, since 2000, http://jthj.ffzg.hr
- curricula
  - University of Zagreb, Faculty of Humanities and Social Sciences
    - Department of Linguistics
      - M.A. study of Linguistics, direction Computational Linguistics
      - educating experts in computational and corpus linguistics
    - Department of Information Sciences
      - a range of courses in NLP
  - University of Zadar
    - Department of Linguistics
      - M.A. study of Linguistics, direction Computational Linguistics

Croatia: LT in industry

- Matica hrvatska & SysPrint: spelling checker, 1997 (MS-Office)
- Novi Liber: online monolin. dictionary, since 2006, http://hjp.srce.hr
- HIDRA: morphologically and multilingually sensitive search-engine for Croatian legislation, 2009, http://cadial.hidra.hr/search.php
- HINA, Croatian News Agency, 2010, http://www.hina.hr
  - automatic classification of newswires
  - automatic keyword and NE extraction and populating metadata
  - using lemmatisation in search engine
- translation and localisation SMEs using M(A)T
  - Integra, http://www.integra.hr
  - Ciklopea, http://www.ciklopea.com
  - Prevoditelj, http://www.prevoditelj.com...
- **historical meeting for LT:** Dubrovnik, 1989
  - *Language Industries: Needs and Perspectives*
  - for the first time experts from CEE met with colleagues from WE
  - J. Sinclair, A. Zampolli, M. Gross / P. Sgall, E. Hajičova, F. Kiefer, J. Bień...

Serbia

Serbia: Research Institutions

- **University of Belgrade**
  - Faculty of Mathematics – language models & tools
  - Faculty of Philology – language resources
  - Faculty of Philosophy – cognitive modelling
  - Faculty of Electrical Engineering – speech
- **Institute Mihajlo Pupin**
  - software tools
- **University of Novi Sad**
  - Faculty of Philosophy – lexicography
  - Faculty of Technical Sciences – speech
- **Serbian Academy of Sciences and Art**
  - Institute for Serbian Language – lexicography
  - Institute for Balkan Studies – multimedia content

Serbia: Language resources and Tools

Resources for Serbian

- Corpus of Contemporary Serbian
- aligned Corpora (TEI, TMX, HTML…)
  - Serbian-English (general & literature)
  - Serbian-French (literature)
  - multilingual (Verne's *Around the World in 80 days*…) & Serbian-Serbian
- morphological e-dictionaries (simple & MWU, proper names)
- Serbian Wordnet & Multilingual database of proper names
- Multimedia ethnographic database

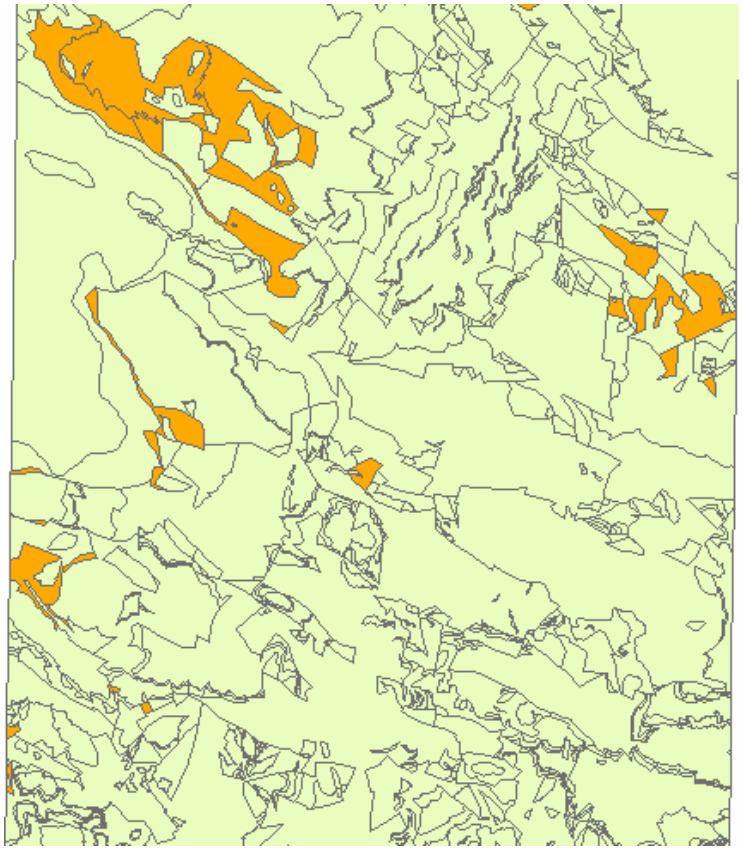
Tools

- Serbian module for **Unitex** (shallow parser, NER) and **NooJ**
- lemmatiser, MSD-tagger (**MulText-East** compliant)
- **LeXimir** (development and interaction between different resources)
- **VebRanka** (multilingual lexically supported query expansion)
- **AlfaNum** (TTS & ASR)

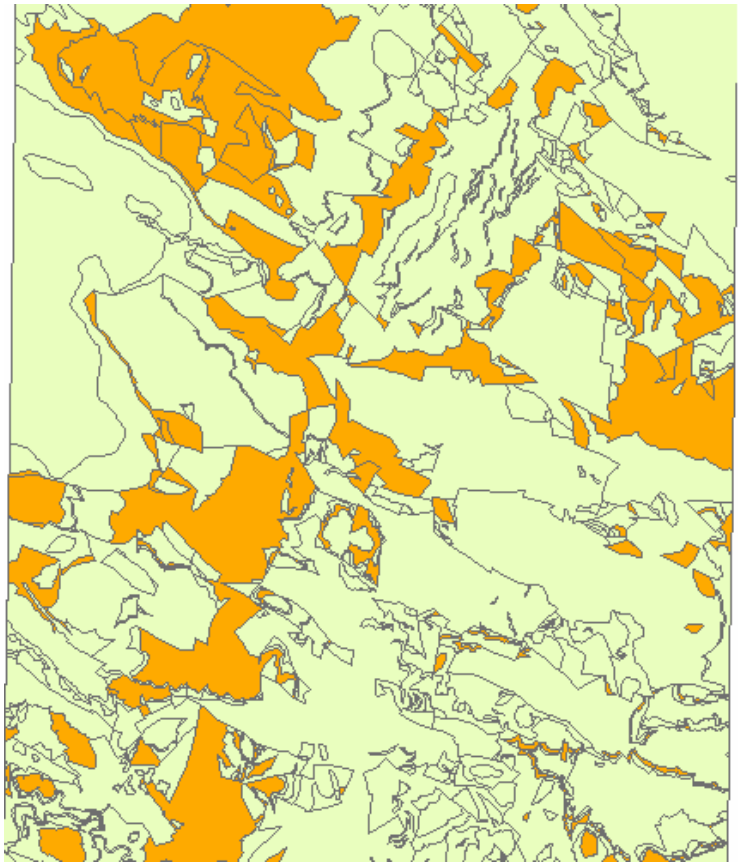
Serbia: A number of applications for a small market

- **IVR systems, call centers, audio logging, etc.**
  - AlfaNum
- **Web monitoring**
  - e-dictionaries + web crawler
- **lexicographic workstation**
  - Serbian Unitex module and resources
- **enformation extraction**
  - Local grammars, lexical resources, named entities
- **organizing digitized content**
  - Wordnet, e-dictionaries, NXD and GIS (ethnographic Serbian material)
- **query expansion for specific domain (e.g. geodata)**
  - Wordnet, e-dictionaries, GIS
- **press clipping**
  - e-dictionaries, named entities extraction
- **Transpoetika – exploring literature on web**
  - e-dictionaries

Language resources for ore retrieval in Serbia



Without LR



With LR

Slovakia

Slovakia: research in LT

- **Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics**
  - today
    - Slovak National Corpus, since 2003, http://korpus.juls.savba.sk
    - Slovak Spoken Corpus, since 2008, http://data.juls.savba.sk/oral
    - parallel corpora (sk-en, sk-cz, fr-sk, ru-sk)
    - Lemmatizer & MSD-tagger
    - Slovak treebank
    - Slovak WordNet
  - projects
    - National: Slovak National Corpus
    - EU: Mondilex, EuroMatrixPlus, Slovak Online
- **Slovak Academy of Sciences, Institute of Informatics**
  - processing of written Slovak, since 2006
  - projects
    - NAZOU, acquisition, organisation and maintenance of knowledge, http://nazou.fiit.stuba.sk
    - Ontea tool for IE and domain dependant metadata generation (incl. language identification and lemmatisation)

Slovakia: research in LT 2

- **Slovak Academy of Sciences, Institute of Informatics, Department of Speech Analysis and Synthesis**
  - acoustic models for telephony speech (SpeechDat-E project)
  - acoustic models for TTS, ASR
- **Slovak Technical University, Department of Telecommunication**
  - speech signal processing in noisy conditions
- **Technical University Košice**
  - voice information retrieval dialogue system for Slovak
  - SAMPA
  - JBOWL (Java Bag-of-Words Library), modular system for NLP comprising tokenization, morphological analysis, lemmatization, disambiguation, syntactic analysis based on ATN networks, clustering and phrase identification, term weighting and indexing
- **University of Žilina, Department of Telecommunications and Multimedia**

Slovakia: LT industry

- Forma s.r.o., http://www.forma.sk
  - spelling-checker (MS-Office)
  - lemmatizer
  - thesaurus
- TEOS Trenčín, http://www.teos.sk/
  - bilingual dictionaries
  - PC Translator, MT system, en-sk
  - Language Teacher, CALL system
- Softec s.r.o., http://www.softec.sk/
  - embedding LT solutions into wider list of IT solutions
- ESET s.r.o., http://www.eset.com/sk/
  - antispam solutions

Conclusions

Conclusions

- **NooJ** development envirnonment, http://www.nooj4nlp.net
  - it will play a significan role in raising the popularity of LT
  - based on the widened concept of local grammars
  - easy to implement and use
  - developed for five CESAR languages already
  - 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
- CESAR is aiming to
  - bring the existing LT for respective languages to the level compatible with other EU languages
  - make the respective LRT accessible also through META-SHARE platform
  - enable cooperation with industrial partners for emerging market of 70+M speakers

Q/A

Thank you for your attention.

http://www.cesar-project.net

office@meta-net.eu
http://www.meta-net.eu
http://www.facebook.com/META.Alliance