Croatian-\(L_x\) Parallel Corpora
Built from the Web
CESAR project in META-NET framework

Marko Tadić, Nikola Ljubešić

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

Slavic Parallel Corpora Workshop, Johannes Gutenberg University
Mainz, Germany, 2012-09-11

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

- META-NET
  - Language Whitepapers
  - META-SHARE platform and Charter
- CESAR project in general
  - geo-linguistic spread, partners in the consortium, general aims
  - NooJ@CESAR
- Croatian parallel corpora within CESAR project
  - hrenWaC, SETimes corpus
  - content analysis
  - usage of these corpora in different tasks
- 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
- META-NET is core of META (Multilingual Europe Technology Alliance)
  - 642 members, 55 countries
- 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 3 ICT-PSP projects

- META-NET (www.meta-net.eu)
  - T4ME: Network of excellence (FP7)
- 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 31 European languages
  - 7 Slavic languages: bg, cz, hr, pl, si, sk, sr
- 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 architecture

- META-SHARE platform for sharing and distribution of LRT
  - open source, whenever possible, some fees applicable for commercial usage
  - www.meta-share.eu
CESAR in META-SHARE

Croatian

15 Language Resources

- Bulgarian-X language Parallel Corpus
- Collocation and Term Extractor
- Croatian Dependency Treebank
- Croatian-English Parallel Corpus
- Croatian-English Parallel Web Corpus
- Croatian Lemmatisation Server
CESAR in META-SHARE

ResourceName: Bulgarian WordNet

Description:
The Bulgarian WordNet was developed by the Department for Computational Linguistics at the Institute for Bulgarian Language, Bulgarian Academy of Sciences, initially within the framework of the BalkaNet project "Multilingual Semantic Network for the Balkan Languages" (IST-2000-29388) and later on...

Read More


MetaShareId: NOT_DEFINED_FOR_V2

Identifier: ELRA-M0041

Availability: available-restrictedUse

Licence: ELRA_END_USER

RestrictionsOfUse: academic-nonCommercialUse

Price: 7114.50
major achievement so far: META-NET Language White Paper Series

- 31 European 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

this snapshot gives a comparative insight where are gaps for particular languages
- directions for LRT to be developed

http://www.cesar-project.net
META-SHARE Charter

- A demand for free usage of digital sources for research purposes
- http://www.cesar-project.net/meta-share/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
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
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
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): Slavic Ls

- **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)

- **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

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

- Serbian

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

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
CESAR Batch 2 (2011-07): Slavic Ls

- **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**

http://www.cesar-project.net
Croatian parallel corpora in CESAR
Building p-corpora: two approaches

- byproduct of a monolingual web corpus
  - crawl the web
  - extract linguistically interesting information
  - identify language
  - try to find document pairs in different languages
  - hrWaC $\rightarrow$ hrenWaC

- focused crawling of known (or suspicious) areas
  - one site / multiple sites
  - pairing documents via regularities in their URL structure
  - http://www.setimes.com
hrenWaC

- hrWaC – the Croatian web corpus
  - 1.2Bw
  - MSD+lemma-tagged
- freely searchable using NoSketchEngine at
  - http://faust.ffzg.hr/bonito2/run.cgi/first_form?corpname=hrwac
- building process of hrenWaC
  - 66.3Mw of English text initially removed from hrWaC
  - 165,194 English vs. 3.4 million Croatian documents – huge search space
- take under consideration only documents inside the same domain
- two types of criteria for identifying pair candidates
  - structural
  - content
hrenWaC

- **structural criteria**
  - identical number of paragraphs
  - cosine similarity of vectors containing numbers of characters per paragraph
    - $\text{LenOfPars}, <243,19,187>$ vs. $<201,15,172>$
  - cosine similarity of vectors containing numbers of sentences
    - $\text{NumOfSents}, <12,1,10>$ vs. $<11,1,10>$

- **content criteria**
  - bilingual Croatian-English dictionary
  - had to lemmatize English side as well
  - $\text{LangOverlap} = |C \cap E| / |C \cup E| \ast \log(|C|)$

- $\text{sim}(d_1,d_2) = \text{LangOverlap} \ast 0.5 + \text{LenOfPars} \ast 0.25 + \text{NumOfSents} \ast 0.25$
Manual inspection

- 12,191 document pairs identified as parallel out of first 33,330 document pairs, accuracy starts at 80%, decreases down to 37%
hrenWaC

- size of hrenWaC
  - 4.3 Mw, 99,001 sentence pairs
- freely downloadable TXT and TMX files at
  http://www.nljubesic.net/resources/corpora/hrenwac/
  http://opus.lingfil.uu.se/hrenWaC.php
  - CC-BY-SA license
- not yet lemmatized and MSD-tagged
- hrenWaC used to train hr-en & en-hr SMT systems in LetsMT! project
  - http://www.letsmt.eu
- future plans with hrenWaC
  - content similarity calculated via
    - probabilistic lexicons
    - sentence alignment software (HunAlign)
  - ongoing crawls for surrounding languages
    - slWaC (~600Mw), srWaC (~700Mw), bsWaC...
SETimes corpus

- Southeast European news portal: http://www.setimes.com
  - “News and views of SouthEast Europe in ten languages – Albanian, Bosnian, Bulgarian, Croatian, English, Greek, Macedonian, Serbian, Romanian, Turkish” (5 Slavic!)

- the first version of the corpus
  - built by Francis M. Tyers and Murat Serdar Alperen
  - presented at LREC2010

- but, serious problems with that version detected
  - de-diacritized text
    - e.g. instead of š and ž only s and z, scattered irregularly through texts
  - portions of English text instead of translation to the other language

- decided to build it from scratch, three-step process
  - recrawling
  - language identification and removing English portions
  - clean-up, solving encoding issues
translations from the English original

**Bosnian**

Internet širi pristup javnosti medijima
Pomoću interneta se sve više širi pristup onlajn medijima u regionu, ali je uslijedila i debata o tačnosti informacija i korisnosti od različitih internet lokacija za korisnike vijesti, kažu eksperti.

**Croatian**

Internet proširuje dostupnost medija javnosti
Iako internet sve više širi dostupnost online medija u regiji, vodi se rasprava o istinitosti informacija i korisnosti različitih internetskih stranica za korisnike vijesti, tvrde stručnjaci.

**Serbian**

Internet povećava pristup javnosti medijima
Internet sve više povećava pristup onlajn medijima u regionu, ali je usledila debata o verodostojnosti informacija i koristi koju novi korisnici imaju od različitih internet lokacija, kažu eksperti.
## SETimes corpus

- amount of Mw after crawling and after clean-up

<table>
<thead>
<tr>
<th>Language</th>
<th>Crawled Mw</th>
<th>Cleaned Mw</th>
<th>&quot;Loss&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albanian</td>
<td>5.2</td>
<td>5.2</td>
<td>0%</td>
</tr>
<tr>
<td>Bosnian</td>
<td>4.7</td>
<td>2.7</td>
<td>42%</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>4.8</td>
<td>4.5</td>
<td>6%</td>
</tr>
<tr>
<td>Croatian</td>
<td>4.5</td>
<td>4.0</td>
<td>11%</td>
</tr>
<tr>
<td>English</td>
<td>4.8</td>
<td>4.8</td>
<td>0%</td>
</tr>
<tr>
<td>Greek</td>
<td>5.2</td>
<td>5.1</td>
<td>2%</td>
</tr>
<tr>
<td>Macedonian</td>
<td>4.8</td>
<td>4.5</td>
<td>6%</td>
</tr>
<tr>
<td>Romanian</td>
<td>5.1</td>
<td>4.7</td>
<td>8%</td>
</tr>
<tr>
<td>Serbian</td>
<td>4.6</td>
<td>4.5</td>
<td>2%</td>
</tr>
<tr>
<td>Turkish</td>
<td>4.0</td>
<td>3.7</td>
<td>8%</td>
</tr>
</tbody>
</table>

http://www.cesar-project.net
SETimes corpus

- SETimes size
  - 43.7 Mw total
  - 2.7-5.2 Mw per language
  - 137k-227 k sentence pairs per language pair
- available for download in TXT and TMX at
  - http://www.nljubesic.net/resources/corpora/setimes/
  - http://opus.lingfil.uu.se/SETIMES.php
  - CC-BY-SA license
- future plans with SETimes
  - new crawls periodically
  - adding topic information (7 general topics) – topic-specific subcorpora
  - distributing not only sentence pairs but whole documents with alignment information
## Content analysis: hrenWaC 1

<table>
<thead>
<tr>
<th>domain name</th>
<th>keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>business</td>
<td>business company service system market development management project</td>
</tr>
<tr>
<td></td>
<td>croatia quality product energy solution process year customer information</td>
</tr>
<tr>
<td></td>
<td>activity croatian</td>
</tr>
<tr>
<td>computing</td>
<td>system data web information site power user network application</td>
</tr>
<tr>
<td></td>
<td>computer time design control service equipment use model device method</td>
</tr>
<tr>
<td>housing / tourism</td>
<td>room century church hotel floor house building terrace apartment</td>
</tr>
<tr>
<td></td>
<td>bathroom kitchen bed town sea street city restaurant wall area</td>
</tr>
<tr>
<td>outdoor / tourism</td>
<td>island sea town area beach tourist park city adriatic place part</td>
</tr>
<tr>
<td></td>
<td>river hotel road boat nature coast water centre</td>
</tr>
<tr>
<td>cuisine / tourism</td>
<td>wine product food oil fish quality cheese production species plant</td>
</tr>
<tr>
<td></td>
<td>tree fruit water forest dish taste meat wood specialty</td>
</tr>
<tr>
<td>domain name</td>
<td>keywords</td>
</tr>
<tr>
<td>------------------------</td>
<td>----------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>regional politics</td>
<td>croatia croatian country republic member force state ministry programme union government nato bosnia war cooperation herzegovina europe slovenia region</td>
</tr>
<tr>
<td>interior politics</td>
<td>party croatian croatia year government hdz sanader president member time company minister state zagreb case sdp office election people</td>
</tr>
<tr>
<td>science &amp; education</td>
<td>croatian student education work science research project development university school croatia study zagreb child knowledge people language life course</td>
</tr>
<tr>
<td>art</td>
<td>music croatian zagreb festival year film art world day event theatre rijeka work city group croatia place time exhibition</td>
</tr>
<tr>
<td>medicine</td>
<td>effect method disease process system water cell body patient temperature mean structure means blood property factor result treatment level</td>
</tr>
<tr>
<td>domain name</td>
<td>keywords</td>
</tr>
<tr>
<td>-------------------</td>
<td>--------------------------------------------------------------------------</td>
</tr>
<tr>
<td>art</td>
<td>world festival year city athens film time event place team university country centre belgrade people greece croatia august town</td>
</tr>
<tr>
<td>Kosovo / region</td>
<td>kosovo country serbia state issue process reform member macedonia law security agreement turkey nato government year membership bih region</td>
</tr>
<tr>
<td>economy</td>
<td>year euros country company cent bank energy government market project investment business economy sector euro development state gas power</td>
</tr>
<tr>
<td>war crimes</td>
<td>operation crime co war bih police herzegovina bosnia serb tribunal court case force hague agency prosecutor year sarajevo serbia</td>
</tr>
<tr>
<td>politics</td>
<td>minister party president election government parliament leader monday vote meeting time kosovo counterpart opposition member deputy wednesday visit coalition</td>
</tr>
</tbody>
</table>
Language discrimination

- Language discrimination = subproblem of language identification
- Between three closely related languages
  - Bosnian, Croatian and Serbian
  - Training set: trilingual corpus (SETimes), 2.7 Mw in each language
- Apply document classification methods for language identification
- Parallel corpus → the difference between documents is not in their content, but language
- Overkill for distant languages, but very efficient for closely related
- Results checked on the test set
  - 600 documents, 200 in each language from
  - http://www.dnevniavaz.ba
  - http://www.vecernji.hr
  - http://www.politika.rs

http://www.cesar-project.net
Analysis of lexical distance

- intersection of lowercased tokens / only lowercase tokens between languages in the trilingual corpus

<table>
<thead>
<tr>
<th></th>
<th>bs</th>
<th>hr</th>
<th>sr</th>
</tr>
</thead>
<tbody>
<tr>
<td>bs</td>
<td></td>
<td>0.952</td>
<td>0.915</td>
</tr>
<tr>
<td>hr</td>
<td>0.950</td>
<td></td>
<td>0.857</td>
</tr>
<tr>
<td>sr</td>
<td>0.930</td>
<td>0.902</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>bs</th>
<th>hr</th>
<th>sr</th>
</tr>
</thead>
<tbody>
<tr>
<td>bs</td>
<td></td>
<td>0.948</td>
<td>0.919</td>
</tr>
<tr>
<td>hr</td>
<td>0.947</td>
<td></td>
<td>0.860</td>
</tr>
<tr>
<td>sr</td>
<td>0.938</td>
<td>0.906</td>
<td></td>
</tr>
</tbody>
</table>
Baselines

- well known tools for the task of language identification tested
  - TextCat 55.5%
    
    |   | bs | hr | sr | acc |
    |---|----|----|----|-----|
    | bs | 90 | 36 | 74 | .45 |
    | hr | 68 | 65 | 67 | .33 |
    | sr | 14 | 8  | 178| .89 |
  
  - Lingua::Identify 48.8%
    
    |   | bs | hr | sr | acc |
    |---|----|----|----|-----|
    | bs | 65 | 117| 18 | .33 |
    | hr | 43 | 151| 6  | .76 |
    | sr | 41 | 82 | 77 | .38 |
  
  - langid.py 87.7%
    
    |   | bs | hr | sr | acc |
    |---|----|----|----|-----|
    | bs | 139| 56 | 5  | .7  |
    | hr | 11 | 187| 2  | .94 |
    | sr | 0  | 0  | 200| 1   |

http://www.cesar-project.net
Naïve Bayes classifier

- model parameters
  - probabilities of words given the category, i.e. language
- the strongest model parameters are actually lexical differences between languages
- accuracy 95.7%

<table>
<thead>
<tr>
<th></th>
<th>bs</th>
<th>hr</th>
<th>sr</th>
<th>acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>bs</td>
<td>181</td>
<td>11</td>
<td>8</td>
<td>.91</td>
</tr>
<tr>
<td>hr</td>
<td>7</td>
<td>193</td>
<td>0</td>
<td>.97</td>
</tr>
<tr>
<td>sr</td>
<td>0</td>
<td>0</td>
<td>200</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Conclusions
Conclusions

- we have presented the
  - META-NET framework
    - incl. META-SHARE and Language Whitepapers series
  - CESAR project
    - incl. NooJ as an open-source development environment
  - results
    - incl. batches 1 & 2
- Croatian-$L_x$ parallel corpora developed within CESAR project
  - hrenWaC, SETimes corpus
  - opportunistic and very cheap corpora of substantial size
  - web is here to grow...
Conclusions and future directions

- example of application of these corpora
  - close languages discrimination
- robust, easily adaptable methods for monitoring the web and collecting and processing what we are interested in
- usage of these parallel corpora in training SMT systems
- usage of comparable corpora
  - detecting parallel sequences from comparable corpora
  - ACCURAT project (finished 2012-06-30)
  - ACCURAT toolkit
Thank you for your attention.

http://www.cesar-project.net

office@meta-net.eu

http://www.cesar-project.net

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