# The CESAR Project: Comprehensive Language Resources and Tools for Europe

**Tamás Váradi**
**coordinator**

**Research Institute for Linguistics, Hungarian Academy of Sciences Budapest, Hungary**
varadi.tamas@nytud.mta.hu

CESAR META-NET Roadshow
Zagreb, 30th November, 2012

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

- The CESAR consortium
- Project objectives
- CESAR in META-SHARE
- Survey of  results
- Gaps and Challenges
- Conclusions

# META-NET & CESAR

# Geo-linguistic position

- CESAR stands for **CE**ntral and **S**outheast Europe**A**n **R**esources
- operates as integral part of META-NET
- 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

# Who is CESAR?

| Participant no. | Participant organisation name | Participant short name | Country |
|---|---|---|---|
| 1 (CO) | Nyelvtudományi Intézet, Magyar Tudományos Akadémia | HASRIL | Hungary |
| 2 | Budapesti Műszaki és Gazdaságtudományi Egyetem | BME-TMIT | Hungary |
| 3 | Sveučilište u Zagrebu, Filozofski Fakultet – University of Zagreb, Faculty of Humanities and Social Sciences | FFZG | Croatia |
| 4 | Instytut Podstaw Informatyki Polskej Akademii Nauk | IPIPAN | Poland |
| 5 | Uniwersytet Lodzki | Ulodz | Poland |
| 6 | Faculty of Mathematics, University of Belgrade | UBG | Serbia |
| 7 | Institut Mihajlo Pupin | IPUP | Serbia |
| 8 | The Institute for Bulgarian Language Prof. Lyubomir Andreychin | IBL | Bulgaria |
| 9 | Jazykovedny Ústav Ludovíta Stúra Slovenskej Akadémie Vied | LSIL | Slovakia |

# The Faces behind CESAR

# Project objectives

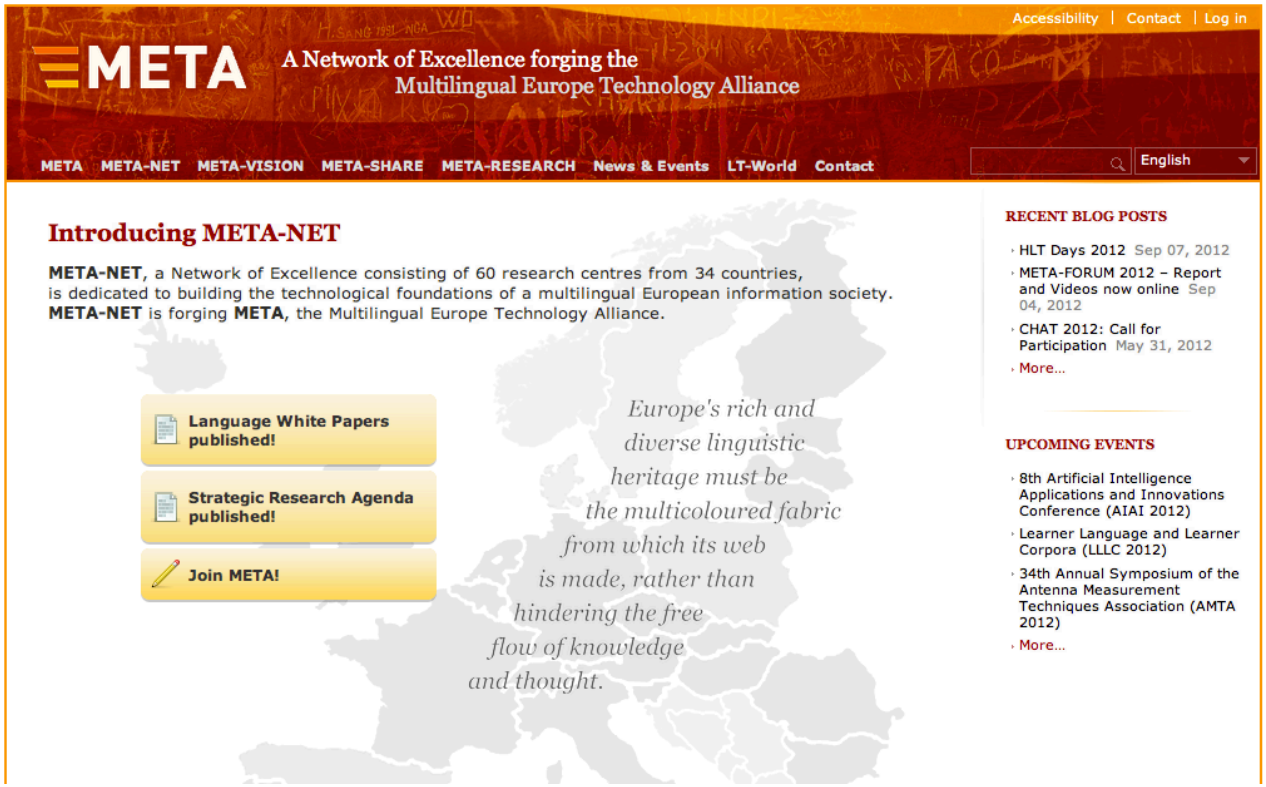
- provide a description of the national landscape in terms of
  - language use, language-savvy products and services, language technologies and resources
- contribute to a pan-European digital language resources exchange (META-SHARE)
  - enhance, extend, document, standardize, cross-link, cross-align resources and tools
- mobilise national and regional stakeholders, public bodies and funding
- reinvigorate cooperation between key technology partners in the region
- collaborate with other partner projects
- bridge the technological gap between this region and the other parts of Europe by

# Timeline

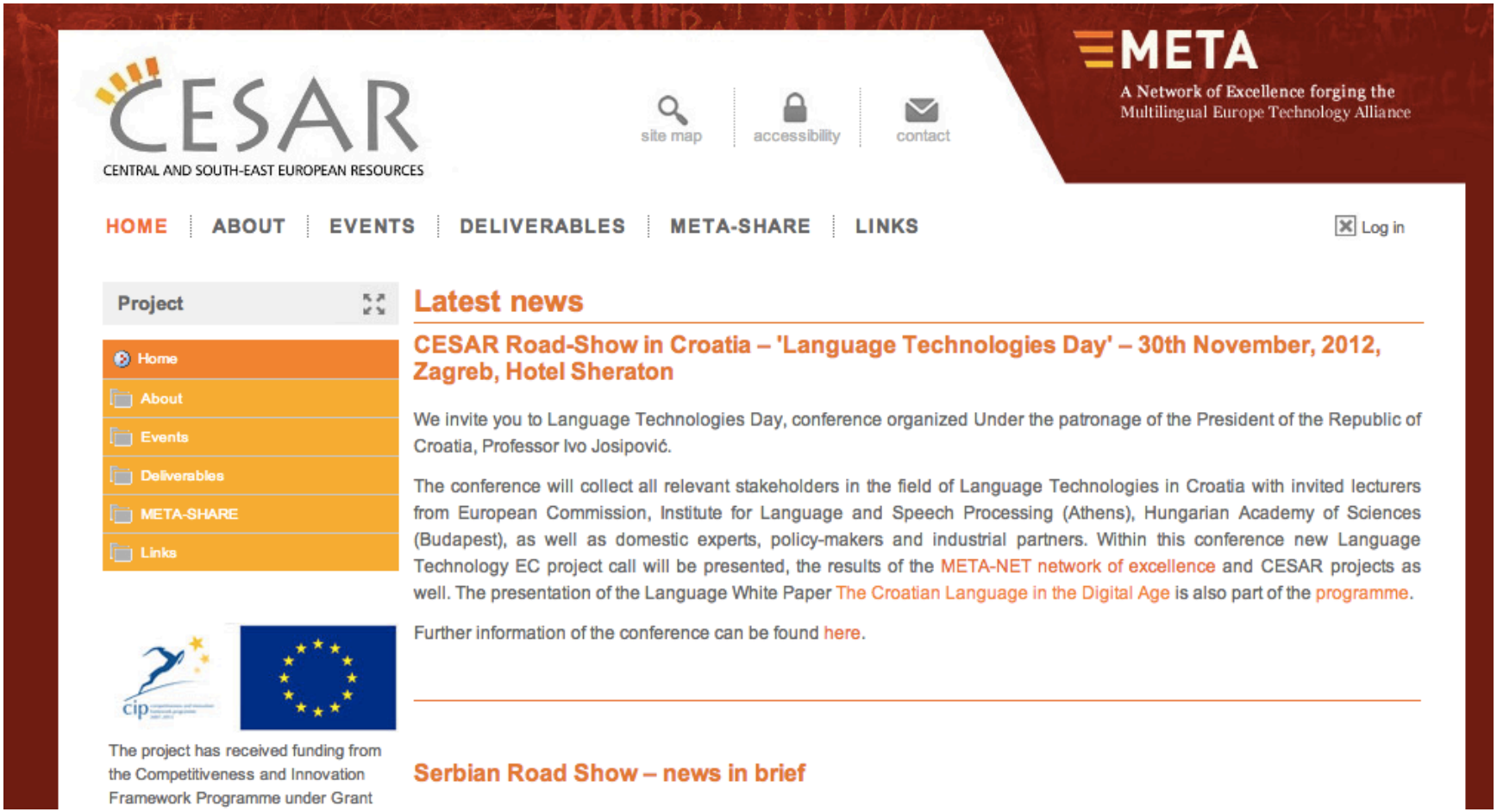
- Project runs between 1st February 2011 and 31st January 2013
- Three major deliverables of resources and tools
- BATCH 1: M10, 30th November 2011
- BATCH2: M18, 31st July 2012
- BATCH3: M24 31st January 2013

# Where to find CESAR

- **www.meta-net.eu**

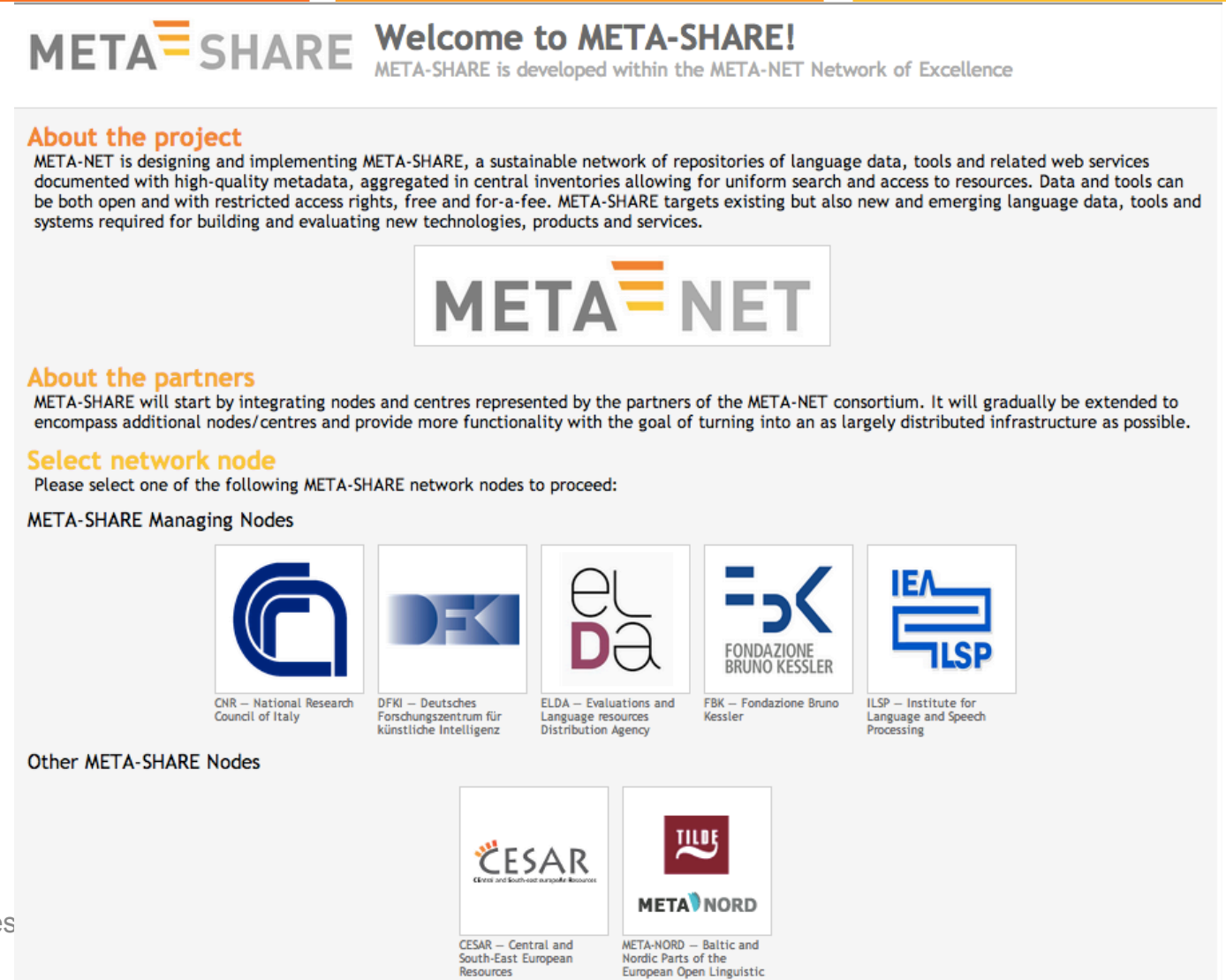


# www.cesar-project.net



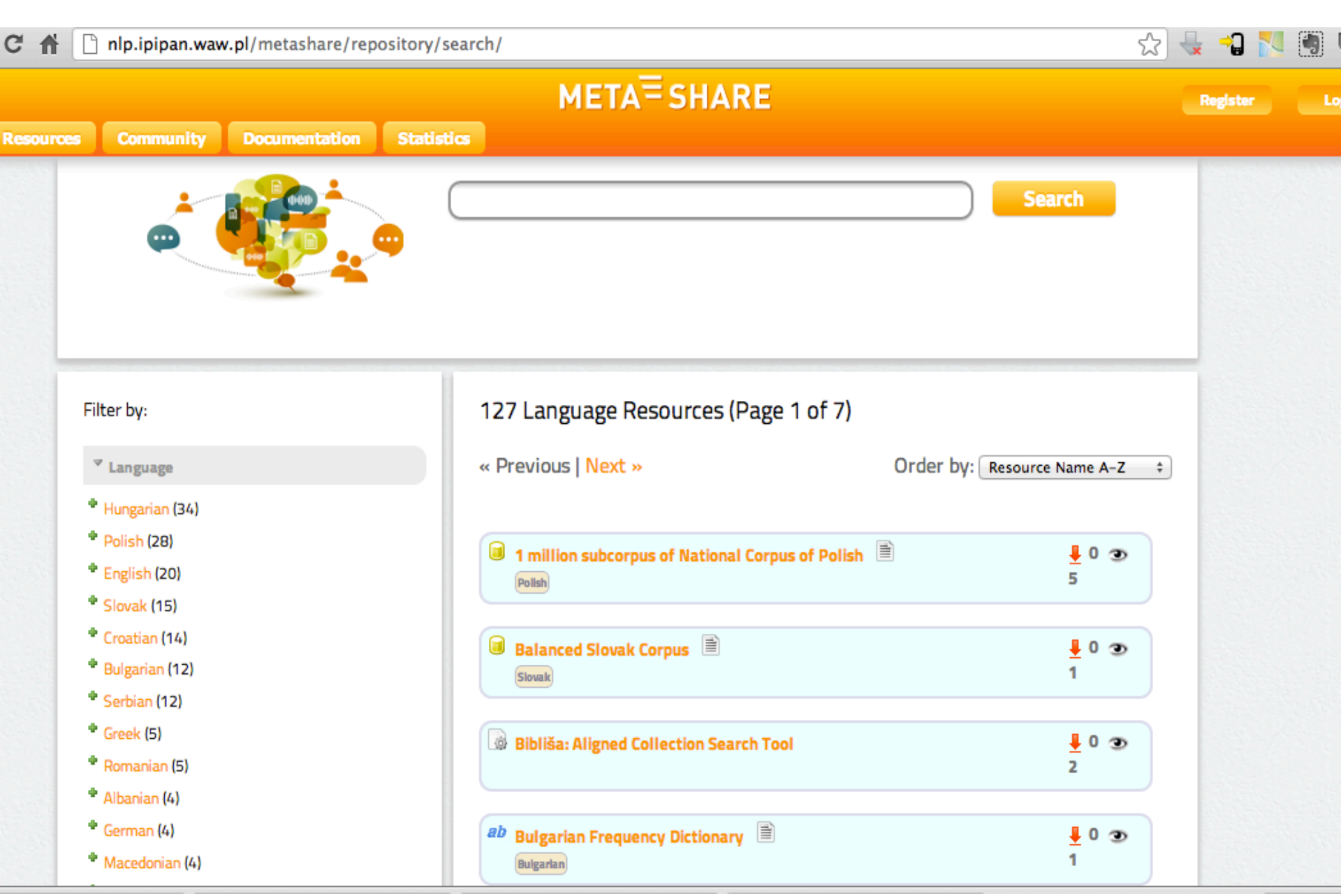
# CESAR in META-SHARE

# www.meta-share.org



# www.cesar-project.net/metashare

nlp.ipipan.waw.pl/metashare
META SHARE
Register
Login
META SHARE
128 language resources at your disposal...
Type in your keywords, please...
Search
Browse
Statistics
What is it? - About the project


nlp.ipipan.waw.pl/metashare/repository/search/
META SHARE
Register
Resources
Community
Documentation
Statistics
Search
Filter by:
Language
Hungarian (34)
Polish (28)
English (20)
Slovak (15)
Croatian (14)
Bulgarian (12)
Serbian (12)
Greek (5)
Romanian (5)
Albanian (4)
German (4)
Macedonian (4)
127 Language Resources (Page 1 of 7)
« Previous | Next »
Order by: Resource Name A-Z
1 million subcorpus of National Corpus of Polish
Polish
0
5
Balanced Slovak Corpus
Slovak
0
1
Bibliša: Aligned Collection Search Tool
0
2
Bulgarian Frequency Dictionary
Bulgarian
0
1

croatian

Search

Filter by:

- Language
- Resource Type
- Media Type
- Availability
- Licence
- Restrictions of Use
- Validated
- Foreseen Use
- Use Is NLP Specific
- Linguality Type
- Multilinguality Type

20 Language Resources

Order by: Resource Name A-Z

**Bulgarian-X language Parallel Corpus** 0 3

Albanian Bosnian Bulgarian Croatian Czech Danish Dutch English Estonian Finish Galician German Greek Hungarian Italian Latvian Lithuanian Macedonian Maltese Polish Portuguese Romanian Slovak Slovenian Spanish Swedish Turkish

**Croatian Dependency Treebank** 0 3

Croatian

Croatian Dependency Treebank is a part of the Croatian National Corpus (i.e. Croatian part of the Croatian-English Parallel Corpus, CW2000) where 4,626 sentences (118,529 tokens) are planned to be manually annotated at the analytical layer following the Prague Dependency Treebank formalism adapted to Croatian. The corpus size is currently 3,465 sentences (88,045 tokens). It is published under CC-BY-NC-SA license.

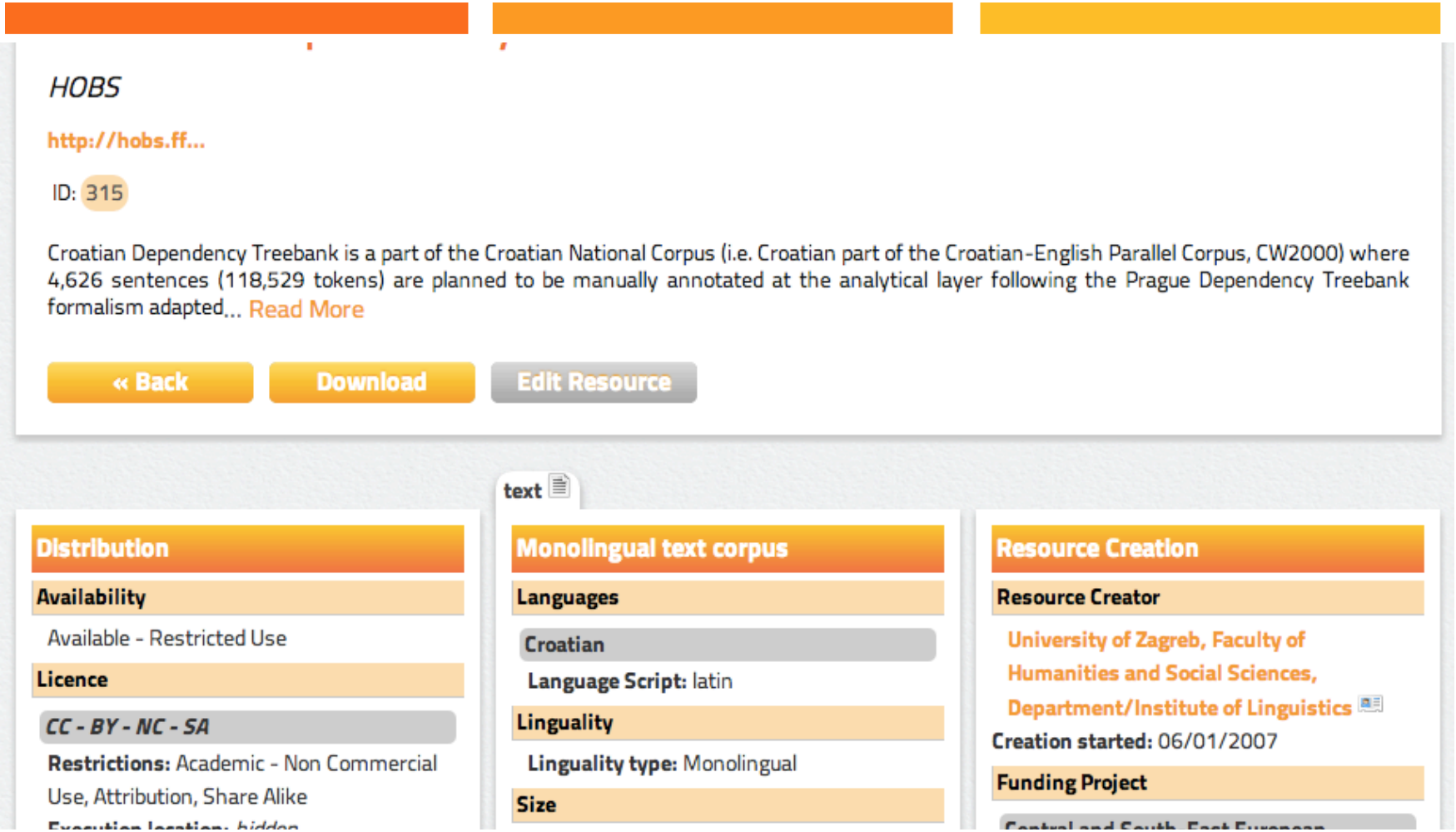
**Croatian Dependency Tre** 0 1

Croatian

**Croatian-English Parallel** 0 1

Croatian English

nlp.ipipan.waw.pl/metashare/repository/browse/croatian-dependency-treebank/e66117a02ccc11e29155000423bfd61c4af8a59f923e43fda9431ab0fa845fa0/


HOBS
http://hobs.ff...
ID: 315
Croatian Dependency Treebank is a part of the Croatian National Corpus (i.e. Croatian part of the Croatian-English Parallel Corpus, CW2000) where 4,626 sentences (118,529 tokens) are planned to be manually annotated at the analytical layer following the Prague Dependency Treebank formalism adapted... Read More
« Back
Download
Edit Resource
text
Distribution
Availability
Available - Restricted Use
Licence
CC - BY - NC - SA
Restrictions: Academic - Non Commercial Use, Attribution, Share Alike
Monolingual text corpus
Languages
Croatian
Language Script: latin
Linguality
Linguality type: Monolingual
Size
Resource Creation
Resource Creator
University of Zagreb, Faculty of Humanities and Social Sciences, Department/Institute of Linguistics
Creation started: 06/01/2007
Funding Project

## Licence Agreement – CC-BY-NC-SA

CREATIVE COMMONS CORPORATION IS NOT A LAW FIRM AND DOES NOT PROVIDE LEGAL SERVICES. DISTRIBUTION OF THIS LICENSE DOES NOT CREATE AN ATTORNEY-CLIENT RELATIONSHIP. CREATIVE COMMONS PROVIDES THIS INFORMATION ON AN "AS-IS" BASIS. CREATIVE COMMONS MAKES NO WARRANTIES REGARDING THE INFORMATION PROVIDED, AND DISCLAIMS LIABILITY FOR DAMAGES RESULTING FROM ITS USE.

***License***

THE WORK (AS DEFINED BELOW) IS PROVIDED UNDER THE TERMS OF THIS CREATIVE COMMONS PUBLIC LICENSE ("CCPL" OR "LICENSE"). THE WORK IS PROTECTED BY COPYRIGHT AND/OR OTHER APPLICABLE LAW. ANY USE OF THE WORK OTHER THAN AS AUTHORIZED UNDER THIS LICENSE OR COPYRIGHT LAW IS PROHIBITED.

BY EXERCISING ANY RIGHTS TO THE WORK PROVIDED HERE, YOU ACCEPT AND AGREE TO BE BOUND BY THE TERMS OF THIS LICENSE. TO THE EXTENT THIS LICENSE MAY BE CONSIDERED TO BE A CONTRACT, THE LICENSOR GRANTS YOU THE RIGHTS CONTAINED HERE IN CONSIDERATION OF YOUR ACCEPTANCE OF SUCH TERMS AND CONDITIONS.

**1. Definitions**

a. "**Adaptation**" means a work based upon the Work, or upon the Work and other pre-existing works, such as a translation, adaptation, derivative work, arrangement of music or other alterations of a literary or artistic work, or phonogram or performance and includes cinematographic adaptations or any other form in which the Work may be recast, transformed, or adapted including in any form recognizably derived from the original, except that a work that constitutes a Collection will not be considered an Adaptation for the purpose of this License. For the avoidance of doubt, where the Work is a musical work, performance or phonogram, the synchronization of the Work in timed-relation with a moving image ("synching") will be considered an Adaptation for the purpose of this License.

b. "**Collection**" means a collection of literary or artistic works, such as encyclopedias and anthologies, or

Please contact the resource maintainer for more information on how to obtain the selected resource under these license

# Results – M18

# CESAR First Batch of Resources

Statistics of resources:

| | HU | | CR | PL | | RS | BG | SK | |
|---|---|---|---|---|---|---|---|---|---|
| | HASRIL | BME-TMIT | FFZG | IPIPAN | ULodz | UBG | IBL | LSIL | |
| **Corpus** | 5 | 5 | 2 | 4 | 3 | 4 | 4 | 4 | **31** |
| **Lexical resource** | 2 | 1 | 2 | 3 | | 1 | 1 | 1 | **11** |
| **Technology, tool, service** | 3 | | 1 | 1 | | 1 | 4 | | **10** |
| | **16** | | **5** | **11** | | **6** | **9** | **5** | **52** |

# CESAR Second Batch of Resources

Statistics of resources:

| | HU | | CR | PL | | RS | BG | SK | |
|---|---|---|---|---|---|---|---|---|---|
| | HASRIL | BME-TMIT | FFZG | IPIPAN | Ulodz | UBG | IBL | LSIL | |
| Corpus | 9 | 2 | 5 | 1 | 1 | 4 | 3 | 7 | **32** |
| Lexical resource | 3 | 0 | 1 | 2 | 2 | 1 | 1 | 3 | **13** |
| Tool, service | 5 | 2 | 3 | 2 | 0 | 0 | 8 | 0 | **20** |
| | **21** | | **9** | **8** | | **5** | **12** | **10** | **65** |

# CESAR Third Batch of Resources

Statistics of resources available for 3rd batch:

| | HU | | CR | PL | | RS | BG | SK | |
|---|---|---|---|---|---|---|---|---|---|
| | HASRIL | BME-TMIT | FFZG | IPIPAN | Ulodz | UBG | IBL | LSIL | |
| Corpus | 4 | 6 | 4 | 4 | 4 | 4 | 1 | - | **27** |
| Lexical resource | 3 | 0 | 1 | 4 | 1 | 1 | 2 | 4 | **16** |
| Tool, service | 2 | 2 | 2 | 7 | 2 | 5 | 7 | 3 | **30** |
| | **17** | | **7** | **22** | | **10** | **10** | **7** | **73** |

# Total resources

| | HU | | CR | PL | | RS | BG | SK | |
|---|---|---|---|---|---|---|---|---|---|
| | HASRIL | BME-TMIT | FFZG | IPIPAN | Ulodz | UBG | IBL | LSIL | |
| Corpus | 18 | 13 | 11 | 9 | 8 | 12 | 8 | 11 | **90** |
| Lexical resource | 8 | 1 | 4 | 9 | 3 | 3 | 4 | 8 | **40** |
| Tool, service | 10 | 4 | 6 | 10 | 2 | 6 | 19 | 3 | **60** |
| | **54** | | **21** | **41** | | **18** | **31** | **22** | **190** |

# ‘In other words – 1st and 2nd batch’

Quick statistics of already submitted LRs:

- monolingual corpus (token) = 1 702 565 806
- paralel corpus (token) = 41 810 000
- record/entry/lexicon = 1 640 579
  - divided between
    - 32 corpora
    - 12 lexical resources
    - 20 tools/services

# Croatian resources in the 1st Batch

- **FFZG**
  - Croatian National Corpus
  - Croatian Morphological Lexicon
  - Croatian-English Parallel Corpus
  - Croatian Lemmatisation Server
  - Croatian Valency Lexicon

# Croatian resources in the 2nd Batch

- FFZG
  - *Croatian National Corpus
  - *Croatian Morphological Lexicon
  - *Croatian-English Parallel Corpus
  - *Croatian Lemmatisation Server
  - *Croatian Valency Lexicon
  - 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

* update of first batch resources

# Proposed Croatian resources in the 3rd Batch

- FFZG
  - Croatian Wordnet
  - Croatian Translations of Acquis
  - Corpus of Narodne novine
  - Croatian-English Giza++ table
  - Croatian Language Corpus
  - Croatian Lemmatisation Web Service
  - Croatian NERC Web Service

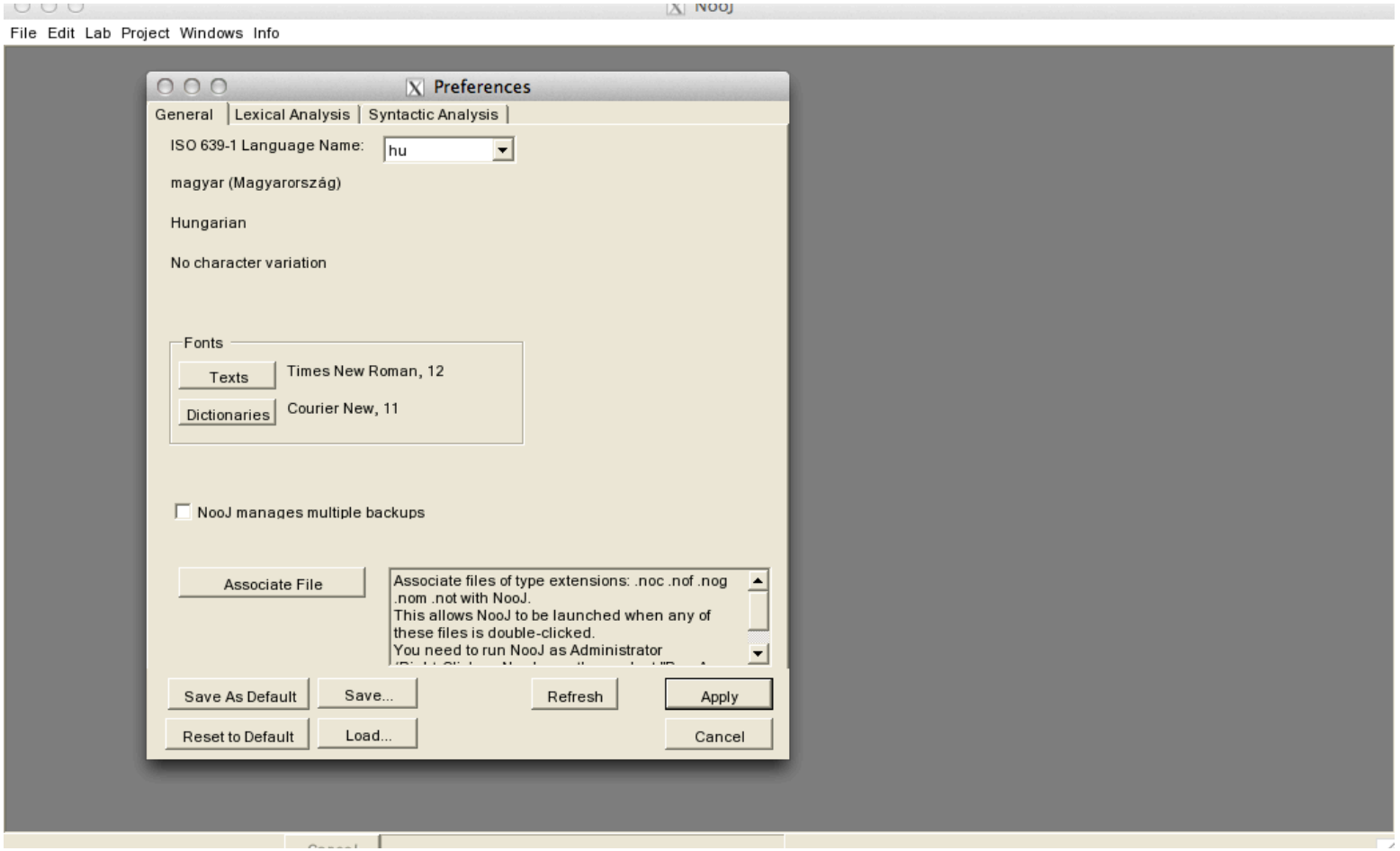
# Distribution of META-SHARE Licence types

| | |
|---|---|
| AGPL | 1 |
| ApacheLicence_V2.0 | 2 |
| BSD-style | 9 |
| CC_BY | 14 |
| CC_BY-NC | 11 |
| CC_BY-NC-ND | 1 |
| CC_BY-NC-S | 2 |
| CC_BY-NC-SA_3. | 2 |
| CC_BY-SA | 12 |
| CLARIN_ACA-NC | 3 |
| CLARIN_RES | 4 |
| ELRA_END_USER | 1 |
| GPL | 12 |
| LGPL | 6 |
| LGPLv3 | 2 |
| MSCommons_BY-NC-SA | 1 |
| MSCommonsCOM-NR-ND-FF | 3 |
| MSCommons_NoCOM-NC-NR | 8 |
| MSCommons_NoCOM-NC-NR | 2 |
| other | 24 |
| proprietary | 6 |
| underNegotiation | 2 |

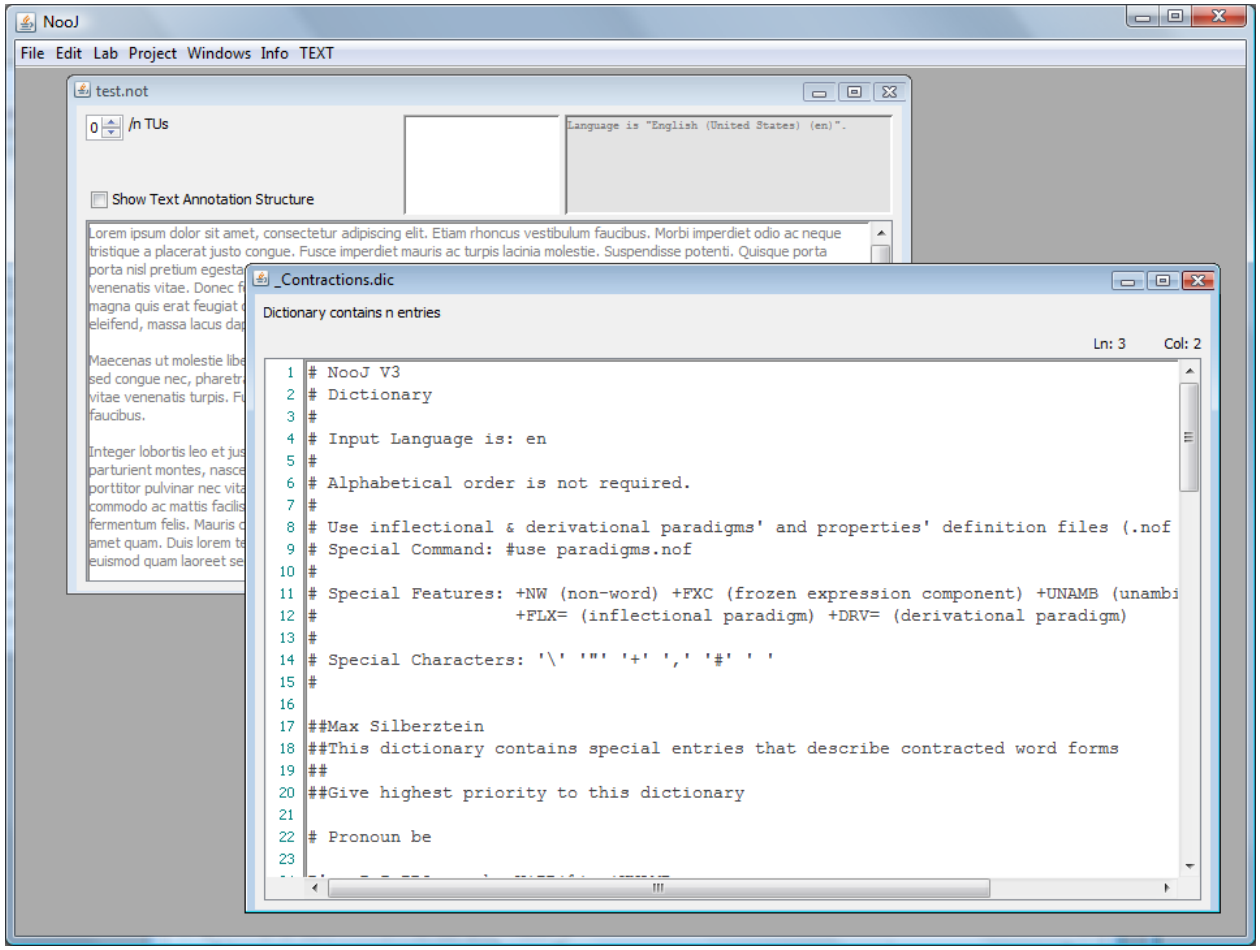
# NooJ

- A linguistic development environment combining fast and robust finite state technology and computational power with ease of use and
- Many CESAR partners had already developed a lot of valuable resources
- Objective: produce open-source and multi-platform version
- Institut Mihajlo Pupin in close collaboration with Max Silberztein, developer of NooJ
- First phase: a version in the MONO system
- Currently, open source  JAVA version in development

# NooJ – Mono version



# NooJ – JAVA version



# Gaps and Challenges*

* Presented at LTC'11, 25-27 November, 2011, Poznan

# Where does CESAR stand?

# MT (top) & Speech Processing (bottom)

META-NET

| excellent | good | moderate | fragmentary | weak or no support |
|---|---|---|---|---|
| | English | French, Spanish | Catalan, Dutch, German, Hungarian, Italian, Polish, Romanian | Basque, Bulgarian, Croatian, Czech, Danish, Estonian, Finnish, Galician, Greek, Icelandic, Irish, Latvian, Lithuanian, Maltese, Norwegian, Portuguese, Serbian, Slovak, Slovene, Swedish |

| excellent | good | moderate | fragmentary | weak or no support |
|---|---|---|---|---|
| | English | Czech, Dutch, Finnish, French, German, Italian, Portuguese, Spanish | Basque, Bulgarian, Catalan, Danish, Estonian, Galician, Greek, Hungarian, Irish, Norwegian, Polish, Serbian, Slovak, Slovene, Swedish | Croatian, Icelandic, Latvian, Lithuanian, Maltese, Romanian |

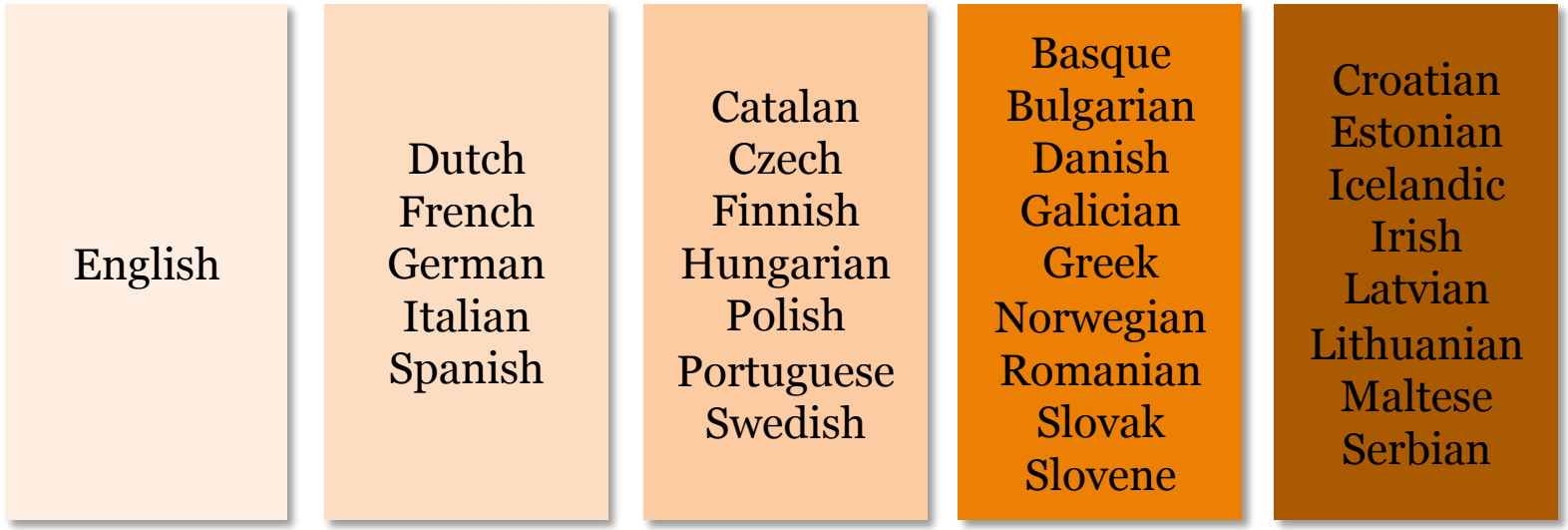
# Text Analysis (top) & Resources (bottom)

META-NET

| excellent | good | moderate | fragmentary | weak or no support |
|---|---|---|---|---|
| | English | Dutch, French, German, Italian, Spanish | Basque, Bulgarian, Catalan, Czech, Danish, Finnish, Galician, Greek, Hungarian, Norwegian, Polish, Portuguese, Romanian, Slovak, Slovene, Swedish | Croatian, Estonian, Icelandic, Irish, Latvian, Lithuanian, Maltese, Serbian |

| excellent | good | moderate | fragmentary | weak/no support |
|---|---|---|---|---|
| | English | Czech, Dutch, French, German, Hungarian, Italian, Polish, Spanish, Swedish | Basque, Bulgarian, Catalan, Croatian, Danish, Estonian, Finnish, Galician, Greek, Norwegian, Portuguese, Romanian, Serbian, Slovak, Slovene | Icelandic, Irish, Latvian, Lithuanian, Maltese |

# Europe's Languages and LT



# Results for language resources

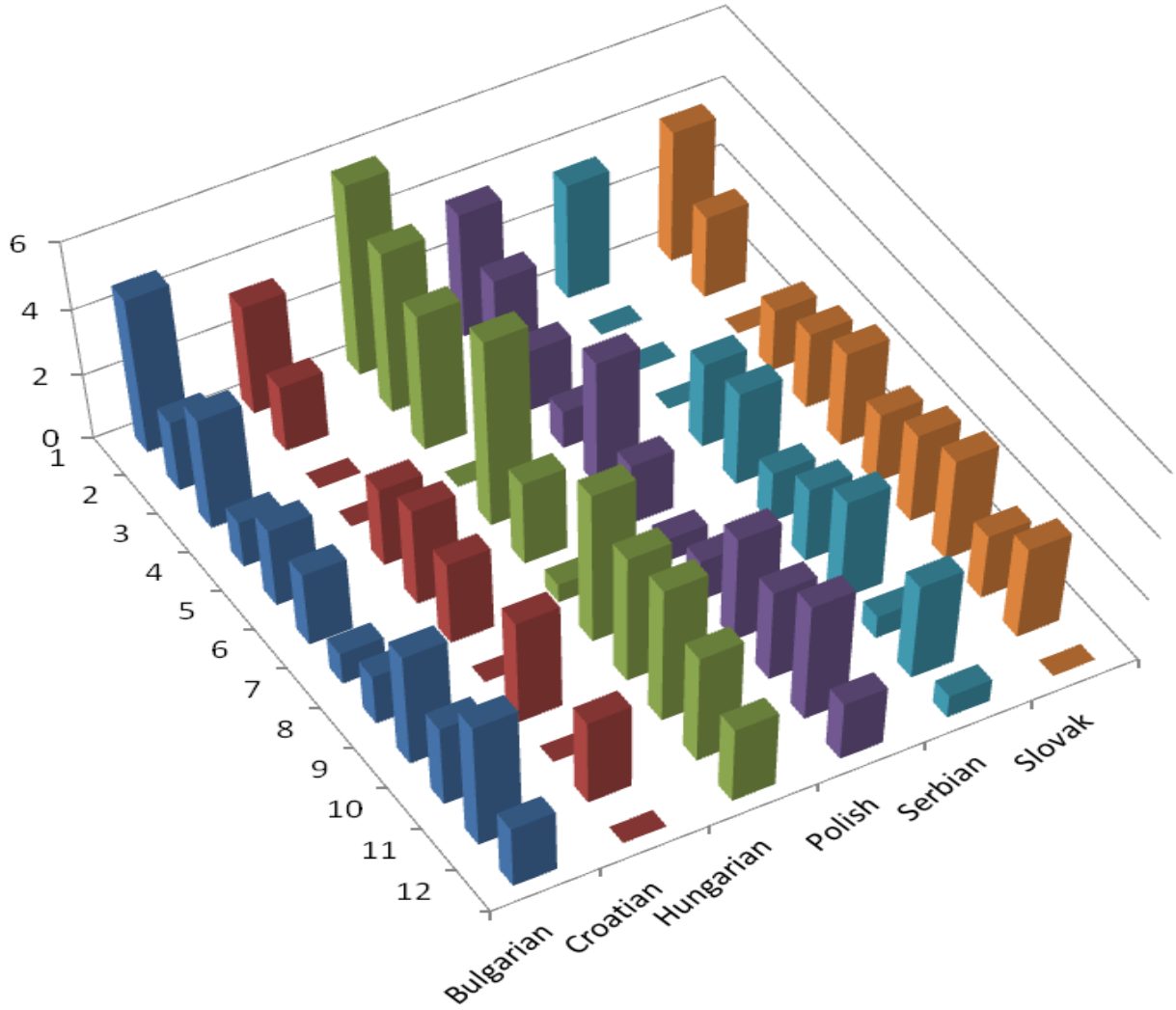
| CESAR languages resources | Bulgarian | Croatian | Hungarian | Polish | Serbian | Slovak | Overall average |
|---|---|---|---|---|---|---|---|
| 1. Reference Corpora | 4.714 | 3.286 | 5.714 | 3.714 | 3.429 | 3.857 | 4.119 |
| 2. Syntax-Corpora (treebanks. dependency banks) | 2.143 | 2.000 | 4.857 | 2.857 | 0.000 | 2.429 | 2.381 |
| **3. Semantics-Corpora** | 3.429 | 0.000 | 4.143 | 1.857 | 0.000 | 0.000 | 1.572 |
| **4. Discourse-Corpora** | 1.429 | 0.000 | 0.000 | 1.143 | 0.000 | 1.857 | 0.738 |
| 5. Parallel Corpora. Translation Memories | 2.429 | 2.429 | 5.714 | 3.857 | 2.571 | 2.286 | 3.214 |
| 6. Speech-Corpora (raw speech data. labelled/annotated speech data. speech dialogue data) | 2.286 | 3.000 | 2.571 | 1.857 | 2.857 | 2.857 | 2.571 |
| **7. Multimedia and multimodal data (text data combined with audio/video)** | 1.000 | 2.571 | 0.571 | 0.714 | 1.571 | 2.143 | 1.428 |
| 8. Language Models | 1.571 | 0.000 | 4.714 | 1.286 | 2.286 | 2.714 | 2.095 |
| 9. Lexicons. Terminologies | 3.571 | 3.286 | 4.000 | 3.286 | 3.143 | 3.143 | 3.404 |
| 10. Grammars | 2.571 | 0.000 | 4.286 | 2.857 | 0.714 | 2.000 | 2.071 |
| 11. Thesauri. WordNets | 4.000 | 2.714 | 3.429 | 3.714 | 3.000 | 2.857 | 3.286 |
| 12. Ontological Resources for World Knowledge (e.g. upper models. Linked Data) | 2.000 | 0.000 | 2.429 | 1.857 | 0.714 | 0.000 | 1.167 |

**below 1.000 in average**; **below 2.000 in average**; **equals 0.000 in cells**

# Discussion

- in half of the categories at least one language has score 0.000 (**50.00%**)
  - under-resourcedness
- two categories where 3 languages have score 0.000
  - 3 Semantics-Corpora
  - 4 Discourse-Corpora

# Results for language resources

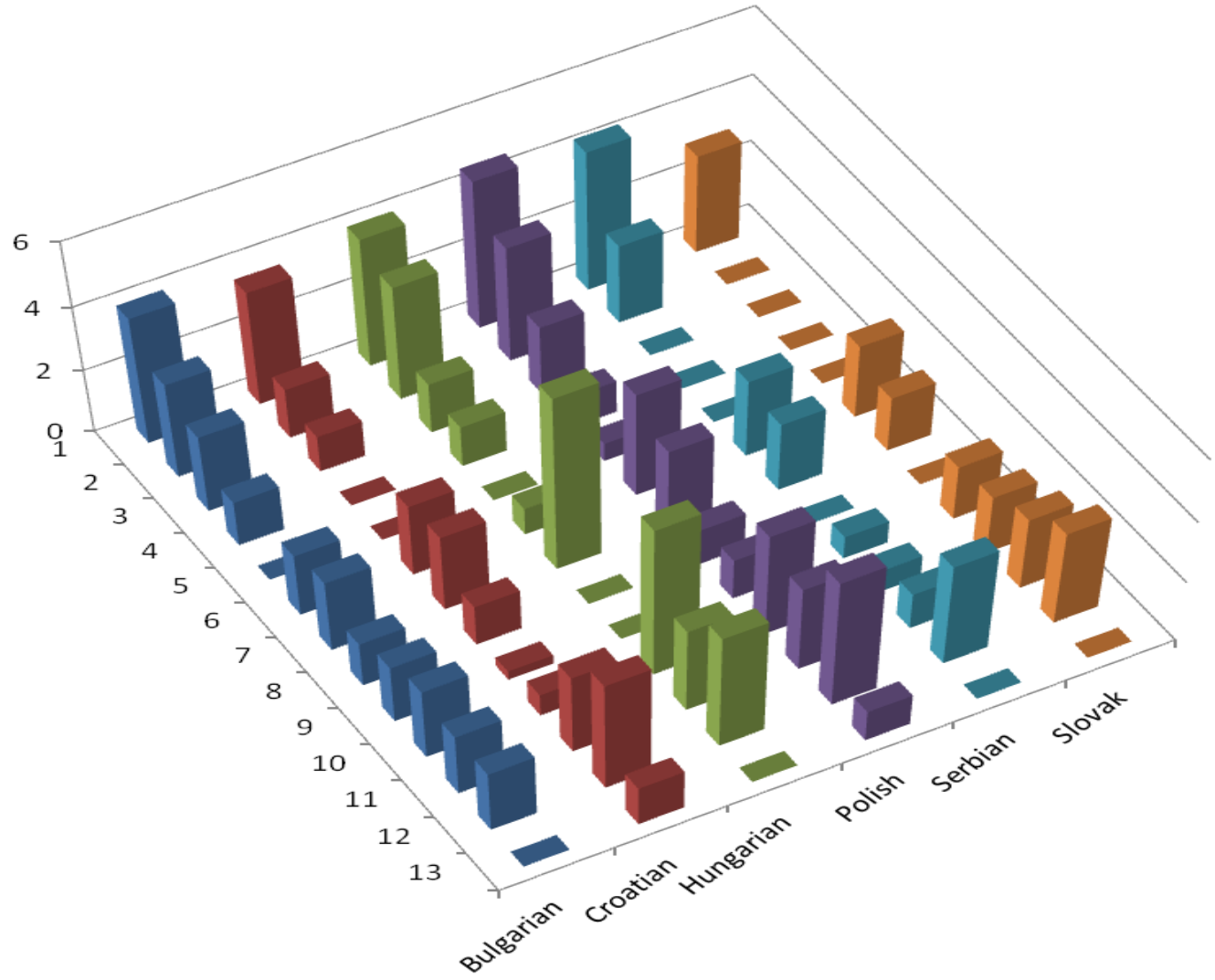


# Results for language tools

| CESAR Language Technology (Tools, Technologies, Applications) | Bulgarian | Croatian | Hungarian | Polish | Serbian | Slovak | Overall average |
|---|---|---|---|---|---|---|---|
| 1. Tokenization. Morphology (tokenization. POS tagging. morphological analysis/generation) | 4.000 | 3.571 | 4.000 | 4.571 | 4.286 | 3.000 | 3.905 |
| 2. Parsing (shallow or deep syntactic analysis) | 3.000 | 1.571 | 3.571 | 3.571 | 2.429 | 0.000 | 2.357 |
| **3. Sentence Semantics (WSD. argument structure. semantic roles)** | 2.429 | 1.143 | 1.571 | 2.143 | 0.000 | 0.000 | **1.214** |
| **4. Text Semantics (coreference resolution. context. pragmatics. inference)** | 1.429 | 0.000 | 1.286 | 1.000 | 0.000 | 0.000 | **0.619** |
| **5. Advanced Discourse Processing (text structure. coherence. rhetorical structure/RST. argumentative zoning. argumentation. text patterns. text types etc.)** | 0.000 | 0.000 | 0.000 | 0.571 | 0.000 | 0.000 | **0.095** |
| 6. Information Retrieval (text indexing. multimedia IR. crosslingual IR) | 2.000 | 2.286 | 0.857 | 3.286 | 2.429 | 2.286 | 2.190 |
| 7. Information Extraction (named entity recognition. event/relation extraction. opinion/sentiment recognition. text mining/analytics) | 2.286 | 2.429 | 5.571 | 2.571 | 2.143 | 1.714 | 2.786 |
| **8. Language Generation (sentence generation. report generation. text generation)** | 1.429 | 1.286 | 0.000 | 1.143 | 0.000 | 0.000 | **0.643** |
| **9. Summarization. Question Answering. advanced Information Access Technologies** | 1.857 | 0.286 | 0.000 | 1.286 | 0.714 | 1.714 | **0.976** |
| 10. Machine Translation | 2.286 | 0.714 | 4.857 | 3.286 | 0.714 | 1.857 | 2.286 |
| 11. Speech Recognition | 2.000 | 2.571 | 2.714 | 2.714 | 1.143 | 2.286 | 2.238 |
| 12. Speech Synthesis | 2.000 | 3.571 | 3.714 | 4.143 | 3.286 | 3.000 | 3.286 |
| **13. Dialogue Management (dialogue capabilities and user modelling)** | 0.000 | 1.286 | 0.000 | 1.000 | 0.000 | 0.000 | **0.381** |

**below 1.000 in average**; **below 2.000 in average**; **equals 0.000 in cells**

# Results for language tools



# Conclusions

- META-NET excellent opportunity
  - to promote LT in Europe
  - to mobilize all stakeholders around a Strategic Research Agenda
  - to create invaluable stock of resources and tools
- CESAR project actively contributing to these aims
- CESAR META-SHARE node
- Language Whitepaper series is a unique instrument to gain a horizontal perspective of the state of the art in various languages
- Croatian resources and tools are valuable components
- There is major work ahead to bridge the technological gap

**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**

# Discussion

- in 5 of 13 categories overall average below 1.000 (**38.46%**)
- in 7 of 13 categories (**53.85%**) at least one language has mark 0.000
  - under-developed tools
- one category where 5 languages have mark 0.000
  - 5 Advanced Discourse Processing
- one category where 4 languages have mark 0.000
  - 13 Dialogue Management
- two categories where 3 languages have mark 0.000
  - 4 Text Semantics
  - 8 Language Generation
- serious under-development regarding tools in CESAR languages