



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
Belgrade, 29th October, 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.



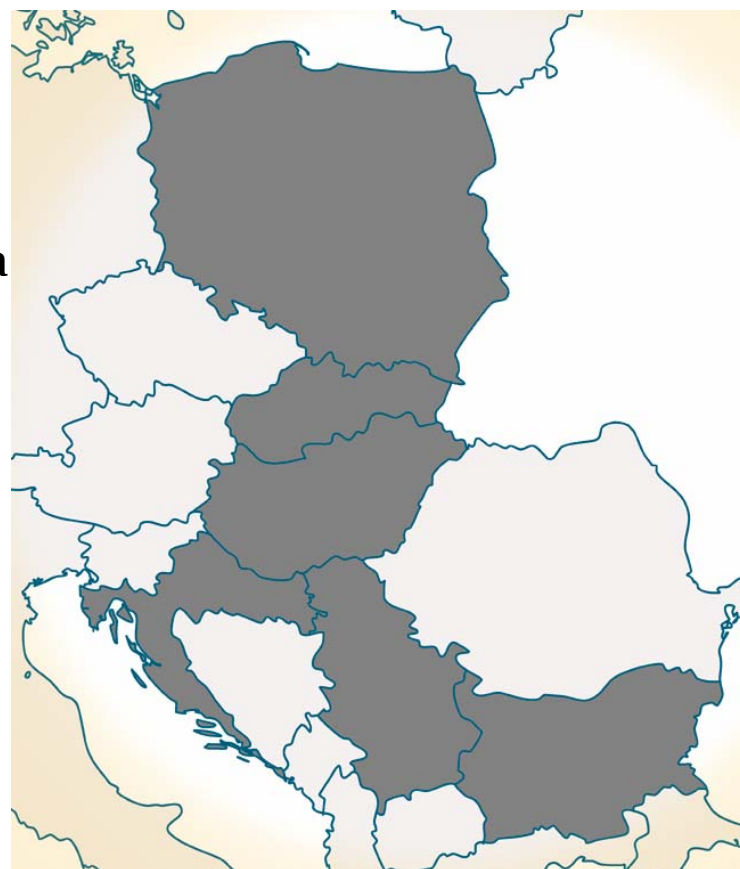
Outline

- ❑ 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 **C**entral 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 Polskiej Akademii Nauk	IPIPAN	Poland
5	Uniwersytet Łódzki	UŁódz	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	Jazykovedný Ústav Ľudovíta 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

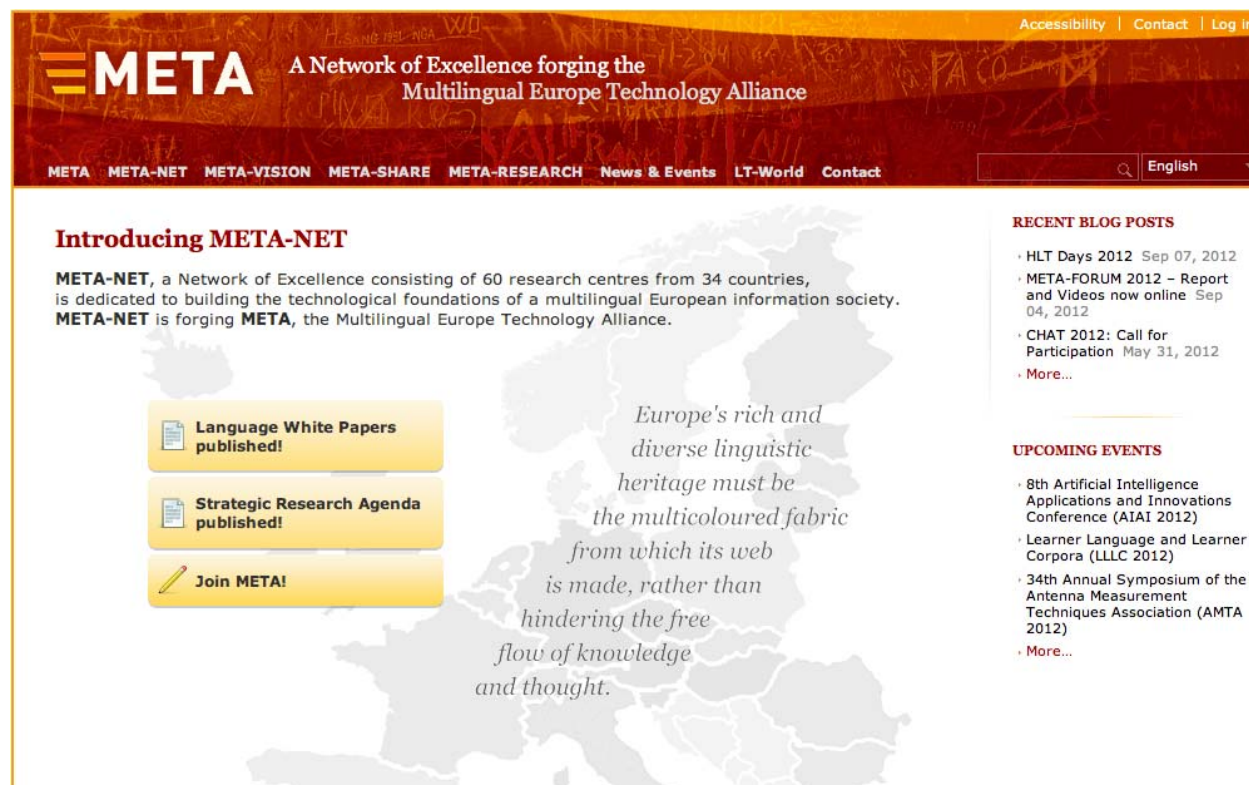
Timeline



- ❑ 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



META A Network of Excellence forging the Multilingual Europe Technology Alliance

Accessibility | Contact | Log in

META META-NET META-VISION META-SHARE META-RESEARCH News & Events LT-World Contact

English

Introducing META-NET

META-NET, a Network of Excellence consisting of 60 research centres from 34 countries, is dedicated to building the technological foundations of a multilingual European information society. **META-NET** is forging **META**, the Multilingual Europe Technology Alliance.

Language White Papers published!

Strategic Research Agenda published!

Join META!

Europe's rich and diverse linguistic heritage must be the multicoloured fabric from which its web is made, rather than hindering the free flow of knowledge and thought.

RECENT BLOG POSTS

- HLT Days 2012 Sep 07, 2012
- META-FORUM 2012 – Report and Videos now online Sep 04, 2012
- CHAT 2012: Call for Participation May 31, 2012
- More...

UPCOMING EVENTS

- 8th Artificial Intelligence Applications and Innovations Conference (AIAI 2012)
- Learner Language and Learner Corpora (LLLC 2012)
- 34th Annual Symposium of the Antenna Measurement Techniques Association (AMTA 2012)
- More...

www.cesar-project.net

CESAR in META-SHARE

www.meta-net.org

META-SHARE Welcome to META-SHARE!

META-SHARE is developed within the META-NET Network of Excellence

About the project

META-NET is designing and implementing META-SHARE, a sustainable network of repositories of language data, tools and related web services documented with high-quality metadata, aggregated in central inventories allowing for uniform search and access to resources. Data and tools can be both open and with restricted access rights, free and for-a-fee. META-SHARE targets existing but also new and emerging language data, tools and systems required for building and evaluating new technologies, products and services.



About the partners

META-SHARE will start by integrating nodes and centres represented by the partners of the META-NET consortium. It will gradually be extended to encompass additional nodes/centres and provide more functionality with the goal of turning into an as largely distributed infrastructure as possible.

Select network node

Please select one of the following META-SHARE network nodes to proceed:

META-SHARE Managing Nodes



CNR — National Research Council of Italy



DFKI — Deutsches Forschungszentrum für künstliche Intelligenz



ELDA — Evaluations and Language resources Distribution Agency



FBK — Fondazione Bruno Kessler



ILSP — Institute for Language and Speech Processing

www.cesar-project.net/metashare

The screenshot shows a web browser window with the address bar displaying www.cesar-project.net/metashare. The website has a dark red header with the CESAR logo (CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES) on the left and the META logo (A Network of Excellence forging the Multilingual Europe Technology Alliance) on the right. Below the header is a navigation bar with links: HOME, ABOUT, EVENTS, DELIVERABLES, META-SHARE (highlighted), and LINKS. A 'Log in' button is also present. The main content area is white. On the left is a sidebar with a 'Project' section containing a menu: Home, About, Events, Deliverables, META-SHARE (highlighted), and Links. The main content area features the META-SHARE logo and the text 'Browse META-SHARE resources related to CESAR'. Below this is the CESAR logo and the text 'CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES'. At the bottom left, there is a logo for the Competitiveness and Innovation Programme (CIP) and the European Union flag, with text stating: 'The project has received funding from the Competitiveness and Innovation Framework Programme under Grant Agreement n° 271022.' At the bottom right, there are 'Send this' and 'Print this' buttons. The footer of the website includes the CESAR logo and navigation links: Home | About | Events | Deliverables | META-SHARE | Links.

<http://www.cesar-project.net>

← → ↻ 🏠 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



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		0
Polish			5
	Balanced Slovak Corpus		0
Slovak			1
	Bibliša: Aligned Collection Search Tool		0
			2
	Bulgarian Frequency Dictionary		0
Bulgarian			1



Filter by:

▼ Language

▣ Serbian (12)

▣ English (4)

▣ Albanian (3)

▣ Bulgarian (3)

▣ Croatian (3)

[more](#)

▼ Resource Type

▼ Media Type

▼ Availability

▼ Licence

▼ Restrictions of Use

12 Language Resources

Order by:



Corpus of Contemporary Serbian



Serbian

↓ 0
2



Corpus of Contemporary Serbian Newspapers and Magazines



Serbian

↓ 0
4



English-Serbian Aligned Corpus



English

Serbian

↓ 0
3



French-Serbian Aligned Corpus



French

Serbian

↓ 0
2



Filter by:

- Language
 - Serbian (12)
 - English (4)
 - Albanian (3)
 - Bulgarian (3)
 - Croatian (3)
 - more
- Resource Type
- Media Type
- Availability
- Licence
- Restrictions of Use
- Validated
- Foreseen Use
- Use is NLP Specific
- Linguality Type
- Multilinguality Type

12 Language Resources

Order by: Resource Name A-Z

- Corpus of Contemporary Serbian** Serbian 0 2
- Corpus of Contemporary Serbian** Serbian
- English-Serbian Aligned Corpus** English Serbian
- French-Serbian Aligned Corpus** French Serbian
- Morphosyntactically tagged Serbian version of Jules Verne's novel "Around the world in 80 days"** Serbian 0 2
- Multilingual Edition of Verne's Novel "Around the World in 80 Days"** Albanian Bulgarian Chinese Croatian Dutch English French German Greek Hungarian Italian Macedonian Polish Portuguese Serbian Slovak Slovenian Spanish 1 2

The Corpus of contemporary Serbian, SrpKor, consists of 4,925 texts. Total size of SrpKor is 118,767,279 words. It is lemmatized and PoS tagged using TreeTagger. SrpKor texts consist of: fiction written by Serbian authors in 20th and 21th century (10,191,092 words), various scientific texts from various domains (both humanities and sciences) (3,542,169 words), legislative texts (6,874,318 words) and general texts (98,159,700 words). General texts represent daily news published in newspaper "Politika" 2000-2002 and 2005-2010, texts in journals and magazines 1991-2002 ("Danica", "Ebit", "Ekonomist", "Glasnik", "NIN", "Ilustrovana politika", "Kalibar", "Moje srce", "Mostovi", "Pravoslavlje", "Svet", "Teološki pogledi", "Trn", "Viva", "Republika"), internet portal texts 2011-2012 (Peščanik), TANJUG agency news 1995-96, newspaper feuilletons published in newspapers "Politika" (2001-2003), "Večernje novosti" (2008-2011) and "Danas" (2002-2006).

Distribution

Availability

Available - Restricted Use

Start date: 12/01/2011

Licence

CC - BY - NC

Restrictions: Academic - Non Commercial Use

Fee: no price

Distribution Access/Medium: Accessible Through Interface

Execution location: *hidden*

Licensors:

Duško Vitas 

Distribution rights holders:

Duško Vitas 

IPR Holder

Duško Vitas 

Contact Person

Duško Vitas 

Monolingual text corpus

Languages

Serbian (118,767,279 Words)

Language Script: Latin

Variety: Ekavian (Type: Dialect)
(118,767,279 Words)

Linguality

Linguality type: Monolingual

Size

118,767,279 Words

Character encoding

ISO - 8859 - 1 (118,767,279 Words)

Domains

science (3,542,169 Words)

literature (10,191,092 Words)

law_politics (6,874,318 Words)

general (98,159,700 Words)

Modalities

Written Language

Annotation

Lemmatization

Tagset:

Resource Creation

Resource Creator

Miloš Utvić 

Duško Vitas 

Creation started: 12/01/2011

Funding Project

**Serbian Language and its Resources:
Theory, Description and Applications**

Funding Type: National Funds

Funder: Serbian Ministry of Education and Science

Funding Country: Serbia

Project duration: 01/01/2011 -
12/31/2015

Metadata

Created: 11/17/2011

Last Updated: 07/27/2012

Source: CESAR

Metadata Creator

Miloš Utvić 

Cvetana Krstev 

Version

Version: v2.1

Last Updated: 08/01/2012

Licence Agreement – CC-BY-NC

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 terms.

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

Serbian resources in the 1st Batch

▣ **UBG**

- Serbian WordNet
- Corpus of contemporary Serbian
- Serbian Lemmatized and PoS Annotated Corpus
- French-Serbian Aligned Corpus
- Multilingual Edition of Verne's Novel "Around the World in 80 Days"
- Organizing digitized material

Serbian resources in the 2nd Batch

- ❑ UBG
 - *Serbian WordNet
 - *Corpus of Contemporary Serbian
 - *Serbian Lemmatized and PoS Annotated Corpus
 - *French-Serbian Aligned Corpus
 - *Multilingual Edition of Verne's Novel "Around the World in 80 Days"
 - *Organizing digitized material
 - English-Serbian Aligned Corpus
 - Serbian NooJ module
- * update of first batch resources
- 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

Proposed Serbian resources in the 3rd Batch

▣ UBG

- Multimedia Ebart Archive
- Terminological Database for geology
- Serbian-English Aligned Literary Corpus
- Named entities evaluation corpus for Serbian
- Named entities module for Serbian
- Language model for Serbian
- Web applications (NE extraction from web pages)
- Emotion classification of

Serbian Texts

- "Semantically tagged Corpus of Contemporary Serbian (preliminary version)"
- A web tool for aligned text search

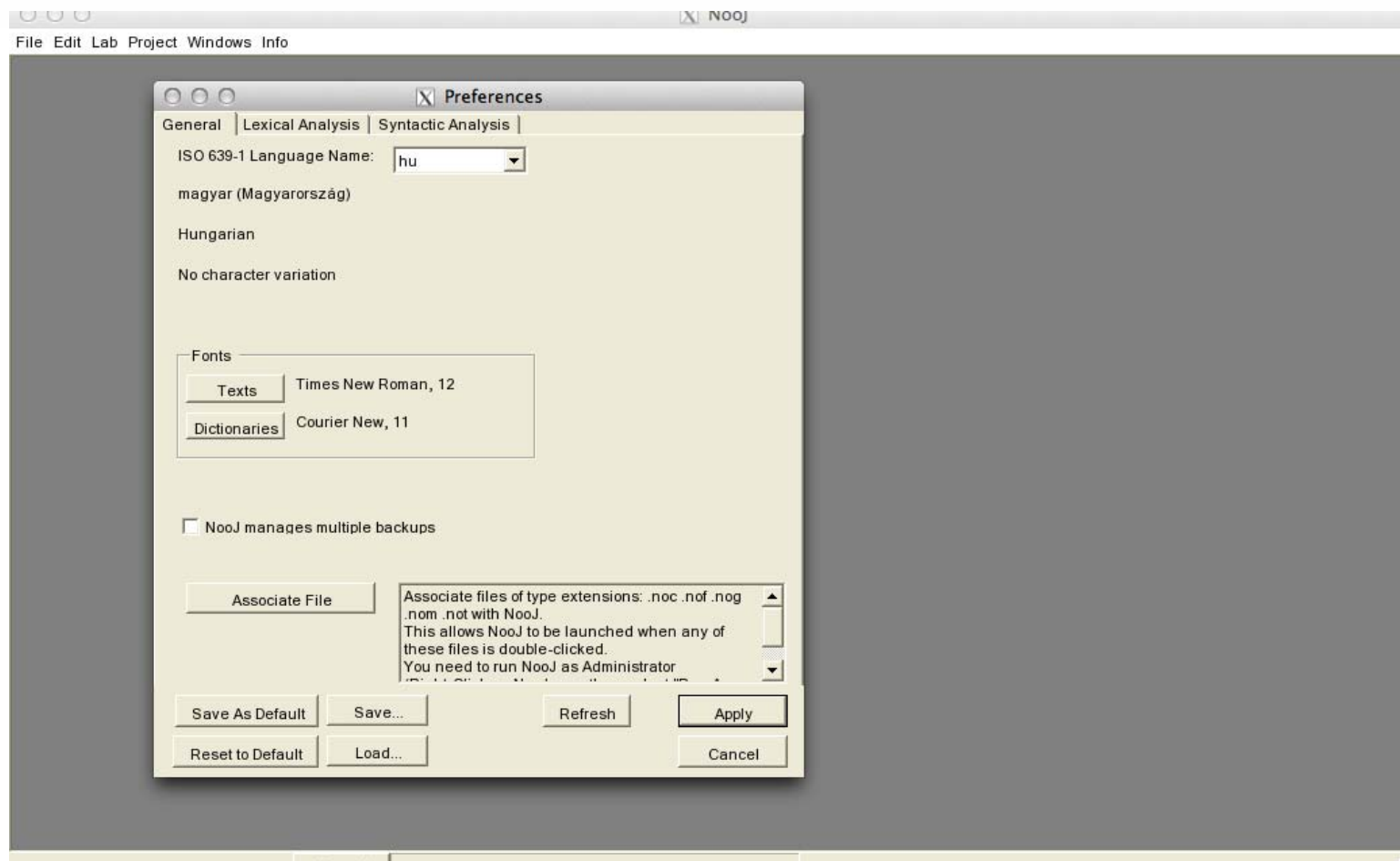
Distribution of META-SHARE Licence types

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

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





Gaps and Challenges*

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

Results for language resources

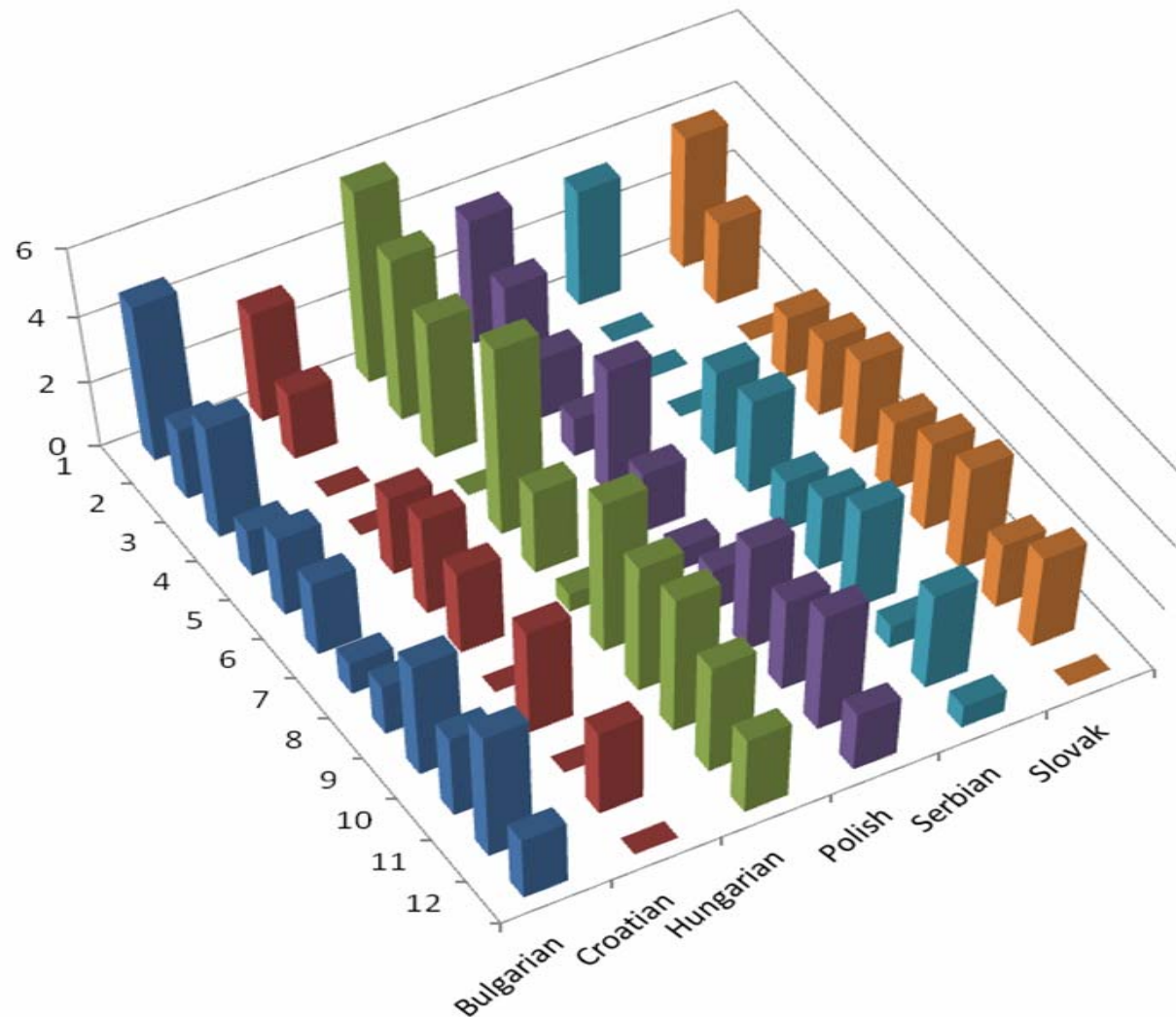
CESAR languages resources	Bulgarian	Croatian	Hungarian	Polish	Serbian	Slovak	Overall average
1. Reference Corpora	4.714	3.286	5.714	3.714	3.429	3.857	4.119
2. Syntax-Corpora (treebanks, dependency banks)	2.143	2.000	4.857	2.857	0.000	2.429	2.381
3. Semantics-Corpora	3.429	0.000	4.143	1.857	0.000	0.000	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 (c.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

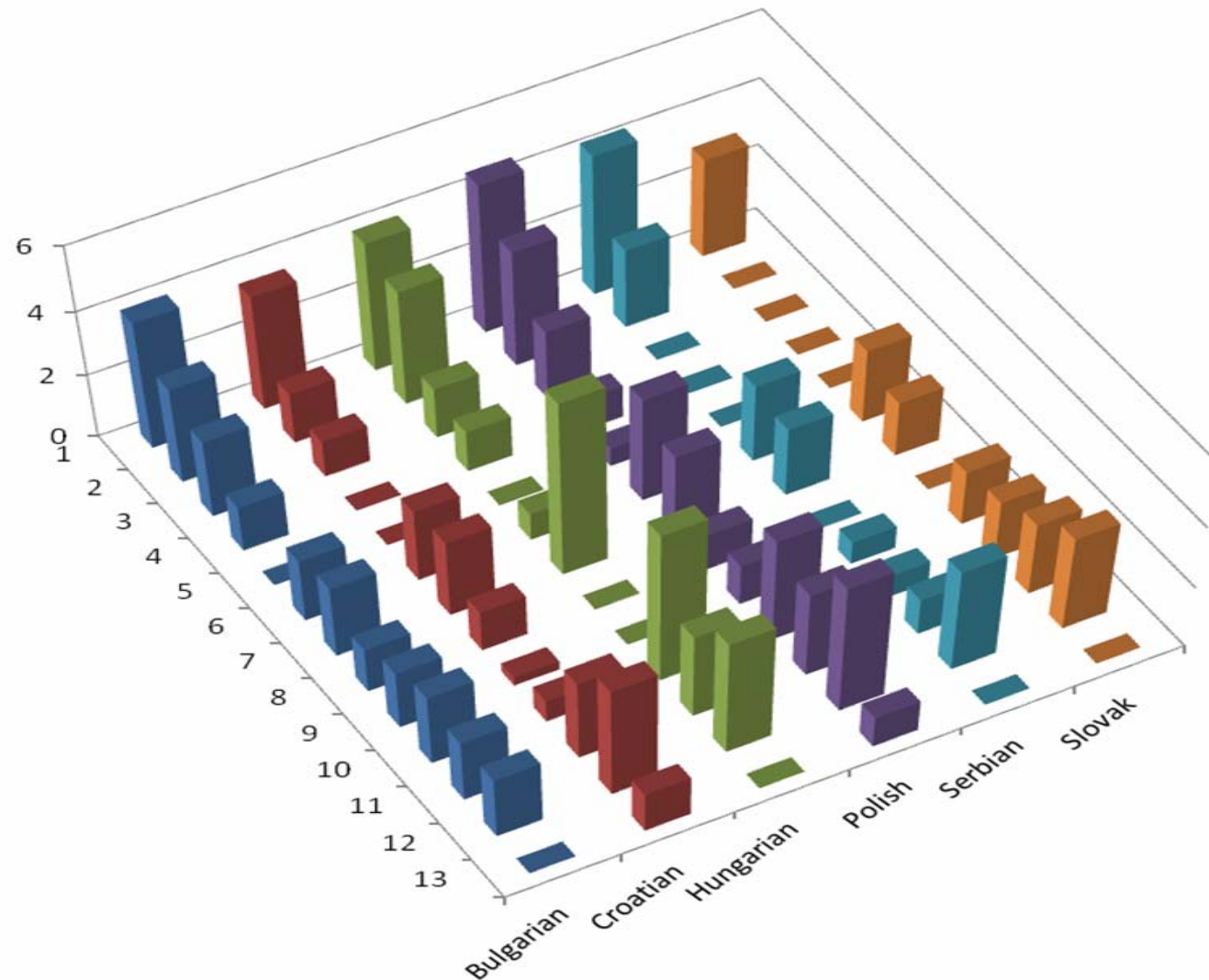


Results for language tools

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



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

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
- ❑ Serbian 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>