Supporting Multilingual Europe
The CESAR initiative

Tamás Váradi

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

CESAR META-NET Roadshow
Sofia, 2nd May, 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

- introducing the CESAR project
- project objectives
- first-year results
- CESAR in META-SHARE
- CESAR in Language Whitepapers
- 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
  - filling obvious and important blind spots in language resources and tools infrastructure

First-year results

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**

Resources – IBL

- Bulgarian morphological dictionary
  - 85000 lemmas, 1 100 000 wordforms
- Bulgarian National Corpus
  - monolingual general corpus
  - fully morpho-syntactically (and partially semantically) annotated
  - about 460,000,000 tokens
  - more than 12,000 samples
- Bulgarian-X language parallel corpora
  - 29 languages
- Bulgarian manually tagged corpora (POS tags, Wordnet senses)
  - manual POS disambiguation of each wordform by language experts
  - more than 200,000 tokens
- Bulgarian Wordnet
  - 40 000 synsets

Actions on resources

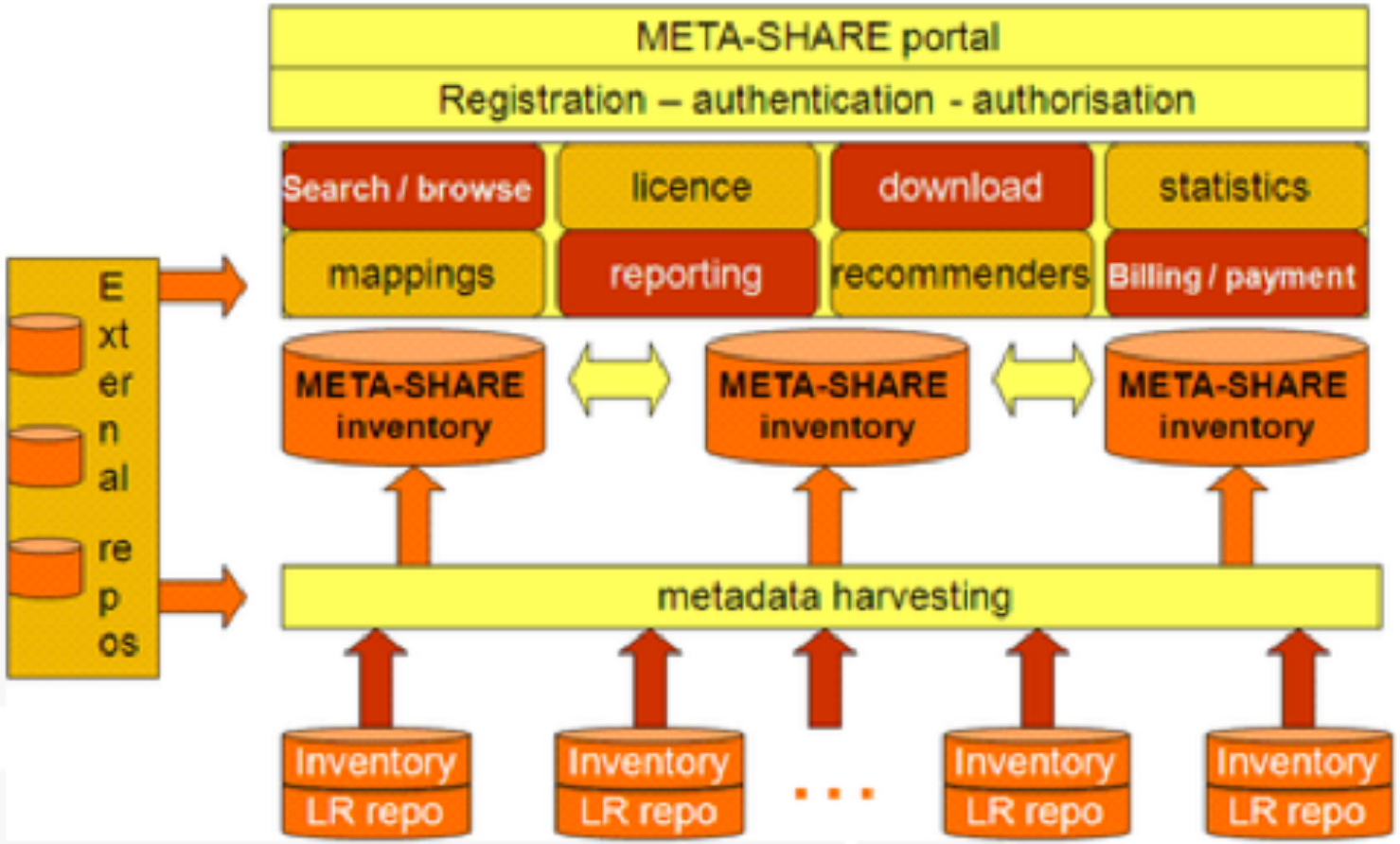
- Bulgarian morphological dictionary
  - online search interface
- Bulgarian National Corpus
  - online search interface
  - web service collocation search
- Bulgarian-X language parallel corpora
  - online search interface
- Bulgarian manually tagged corpora (POS tags, Wordnet senses)
  - online search interface
- Bulgarian Wordnet
  - expansion, consistency checks up, extensive documentation, web access, onling search interface

Tools

- Bulgarian tools
  - Bulgarian Spell Checker for Windows
  - Bulgarian Spell Checker web service
  - WordNet web service
  - Collocation web service

CESAR in META-SHARE

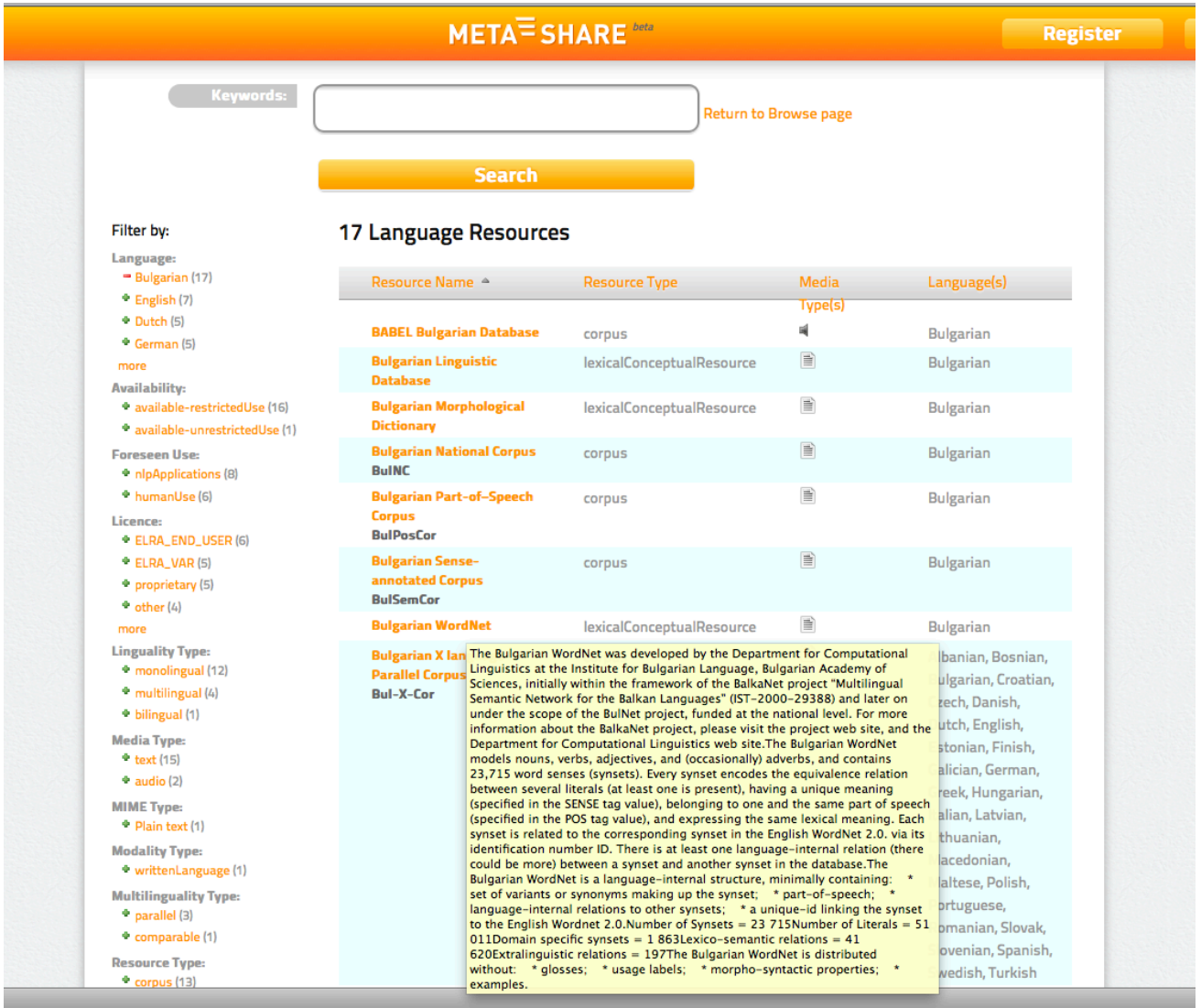
META-SHARE architecture



META-SHARE functions

- user registration, authorization and authentication
- resource/tool description and uploading
- browsing, searching and downloading
- language resources validation, maintenance, evaluation
- language resources archiving, versioning and preservation
- statistics, reporting, recommendations
- access, distribution and referral services including IPR and legal clearance
- billing and payment

CESAR in META-SHARE



CESAR in META-SHARE

META-NET

CESAR
CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES

identificationInfo

ResourceName	Bulgarian WordNet
ResourceName	WordNet bulgare
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
Description	WordNet bulgare a été développé par le Département de linguistique computationnelle de l'Institut pour la Language Bulgare, à l'Académie des sciences bulgare, tout d'abord dans le cadre du projet BalkaNet "Réseau Sémantique multilingue pour les langues des Balkans" (IST-2000-29388), puis plus tard... Read More
Url	http://catalog.elra.info/product_info.php?products_id=802
MetaShareId	NOT_DEFINED_FOR_V2
Identifier	ELRA-M0041

distributionInfo

Availability	available-restrictedUse

licenceInfo

Licence	ELRA_END_USER
RestrictionsOfUse	academic-nonCommercialUse
Price	7114.50

CESAR in Language Whitepapers*

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

META-NET Language Whitepapers

- 1st edition
  - 30 European languages
  - META-FORUM, Budapest, 2011-06
- two tables from 1st edition presenting
  - resources
  - tools/services
- tables with non-merged categories used
  - more detailed list of LR&T categories present
  - allows for more finegrained detection
- used as a data source about CESAR languages for the analysis of
  - level of development of particular area of LR&T
  - characteristic gaps in particular area of LR&T
- here we are interested in gaps

Method for processing scores

- avoiding “vertical” survey of existing LR&T for each language
- favouring “horizontal” approach
  - within the same category
  - for all languages
- method allows insight into situation with
  - groups of languages
  - particular language
- procedure of obtaining scores (led by META-NET for all 30 languages)
  - subjective scores given by national level experts
  - averaged between more experts
- scores and categories
  - scores taken over as they were presented in LWPs
  - list of categories (undergone re-categorisation in 2nd edition of LWP)

Method for processing scores

- for each LR&T category and for each language
  - the average of all marks was calculated
  - no additional weighting for different score dimensions
  - e.g.

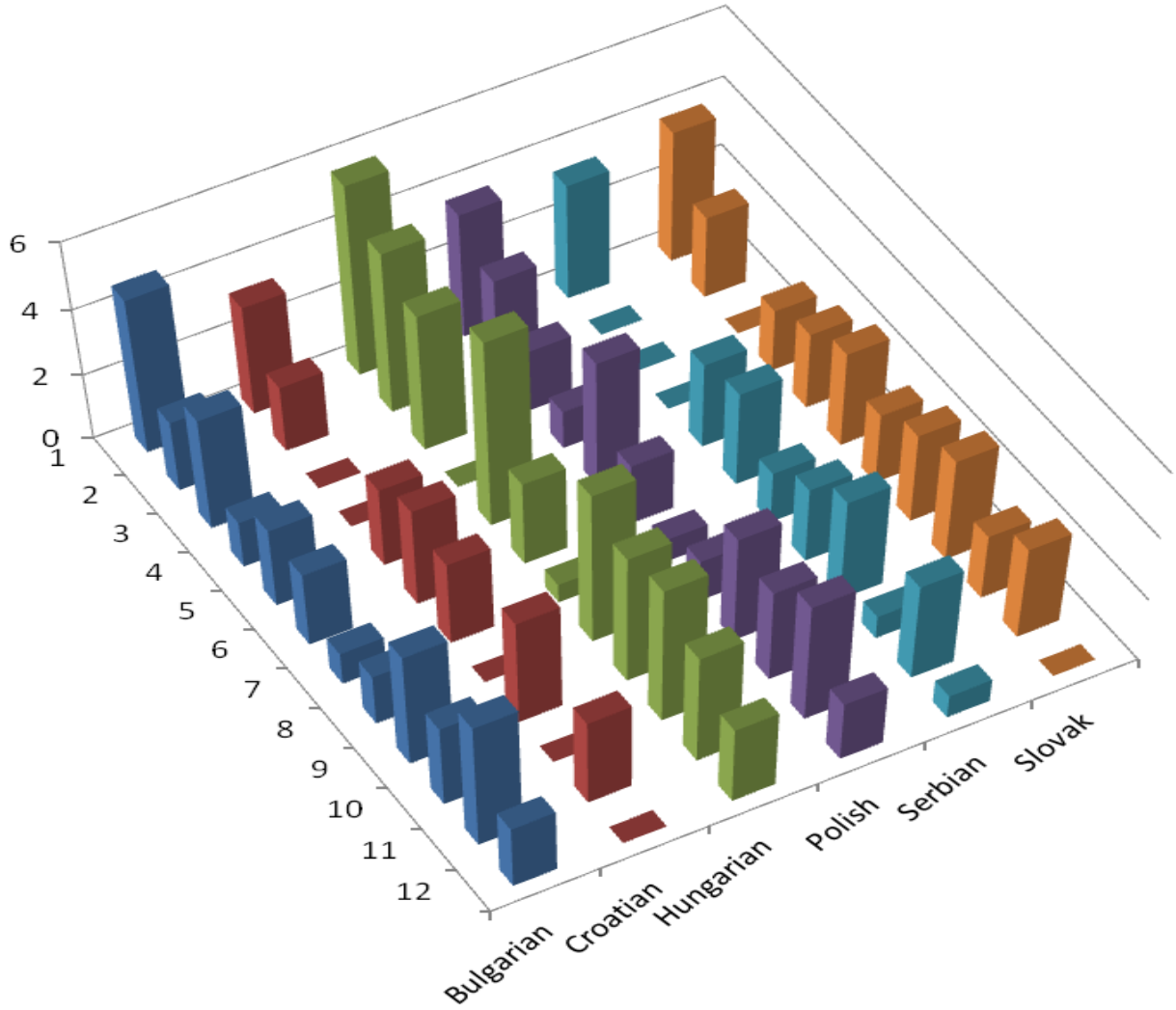
Bulgarian Tools	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability	Average
Parsing (shallow or deep syntactic analysis)	2	2	4	4	3	3	3	3.00

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**

Results for language resources



Discussion

- in half of the categories at least one language has score 0.000 (**50.00%**)
  - under-resourcedness of CESAR languages
- two categories where 3 languages have score 0.000
  - 3 Semantics-Corpora
  - 4 Discourse-Corpora
- also considerable discrepancy between languages in the same category
  - e.g. 3 Semantics-Corpora

bg	3.429	pl	1.857
hu	4.143	sr	0.000
hr	0.000	sk	0.000

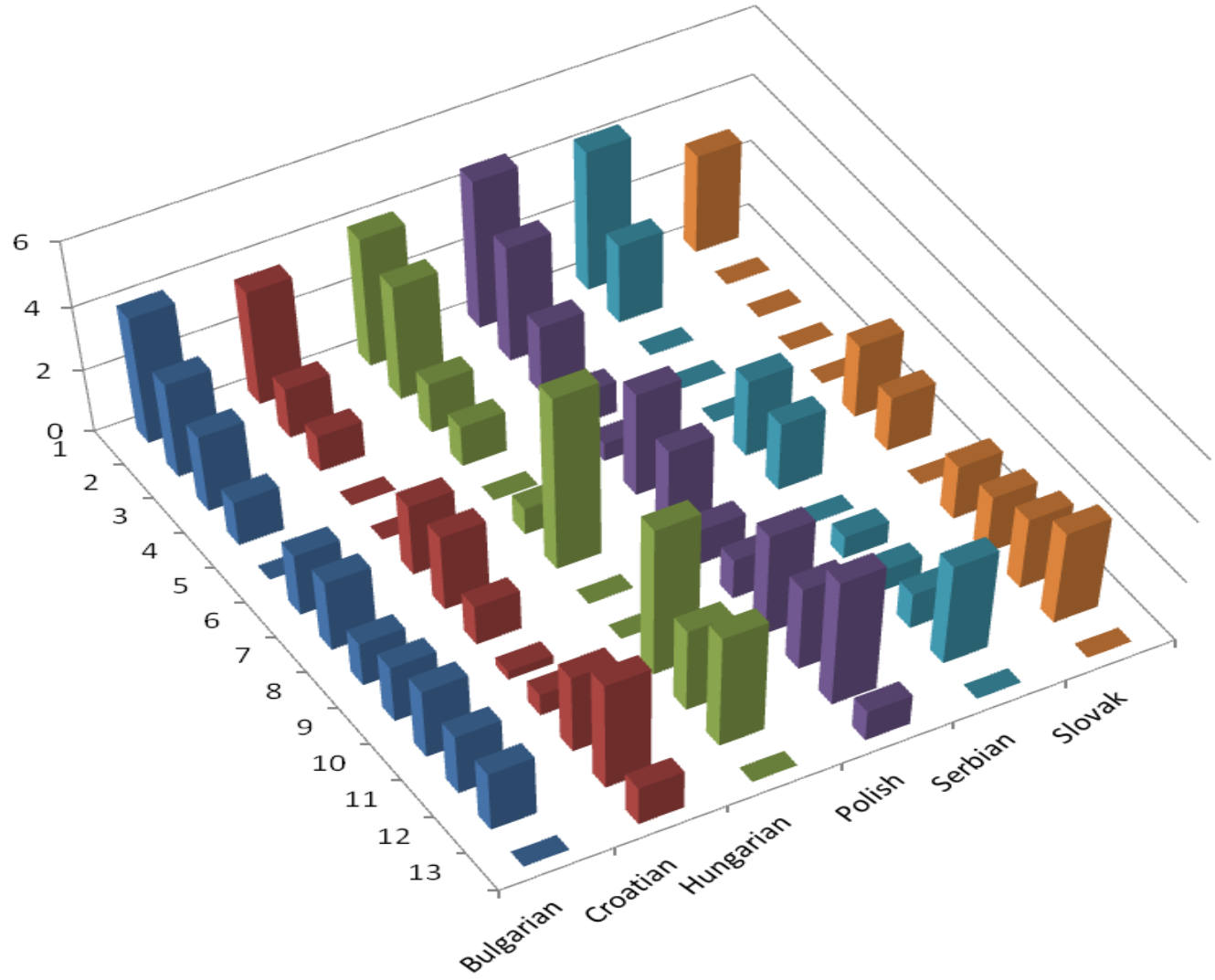
  - category well defined?
  - misunderstanding in criteria for giving scores between national experts?
- similar discrepancy does not appear in other categories
- individual languages snapshot
  - vertical reading of the table

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
- individual languages snapshot
  - vertical reading of the table

Discussion

- preliminary investigation
- harmonized acceptable scores (over **3.000**) in all languages
  - resources (4 of 12 categories)
    - 1 Reference Corpora (**4.119**, range 4.714 – 3.286)
    - 5 Parallel Corpora. Translation Memories (**3.214**, range: 5.714 – 2.286)
    - 9 Lexicons. Terminologies (**3.404**, range: 4.000 – 3.143)
    - 11 Thesauri. Wordnets (**3.286**, range: 4.000 – 2.714)
  - tools (2 of 13 categories)
    - 1 Tokenization. Morphology (**3.905**, range: 4.571 – 3.000)
    - 12 Speech synthesis (**3.286**, range: 4.143 – 2.000)
- serious gaps detected in certain categories
  - for all CESAR languages together or separately
- reccomendation to use these figures
  - targeted development of deficient resources and tools
  - negotiations for support on the national level

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
- Bulgarian resources and tools are valuable components
- there is major work ahead to bridge the technological gap

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