CESAR: Comprehensive Language Resources and Tools for Europe

Tamás Váradi

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

CESAR META-NET Roadshow
Warsaw, 26th September, 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

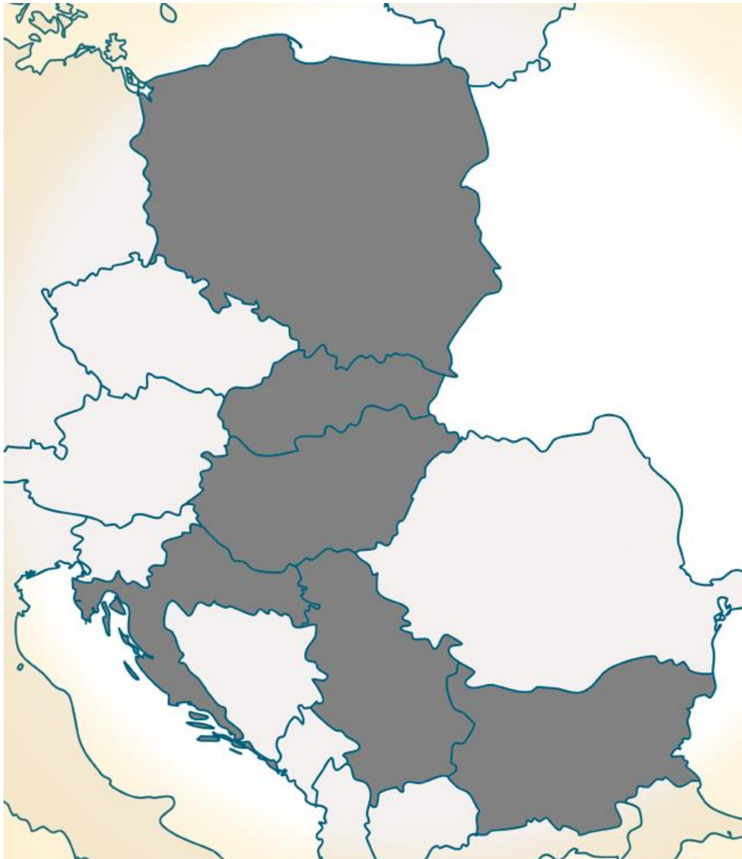
META-NET & CESAR

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

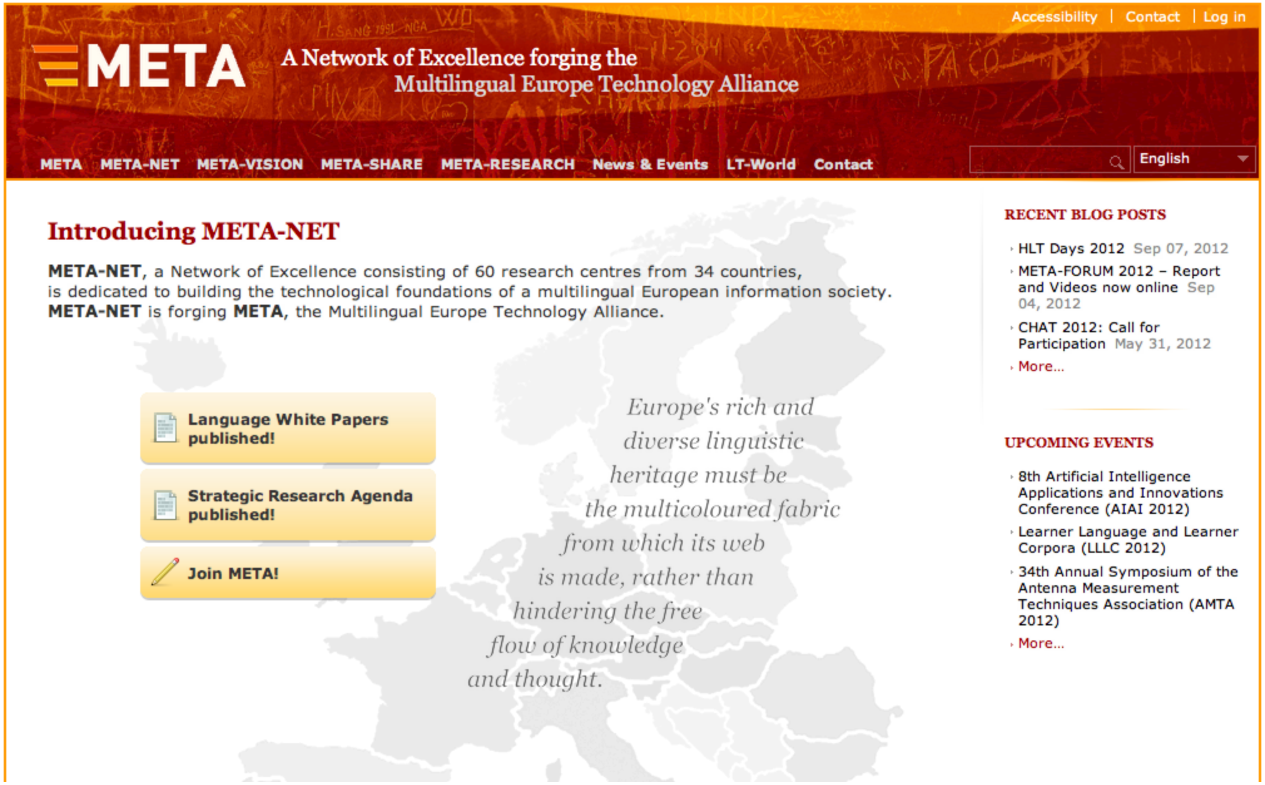
- provide a description of the national landscape in terms of
  - language use, language-savvy products and services, language technologies and resources
- contribute to a pan-European digital language resources exchange (META-SHARE)
  - enhance, extend, document, standardize, cross-link, cross-align 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**



Accessibility | Contact | Log in

META

A Network of Excellence forging the Multilingual Europe Technology Alliance

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

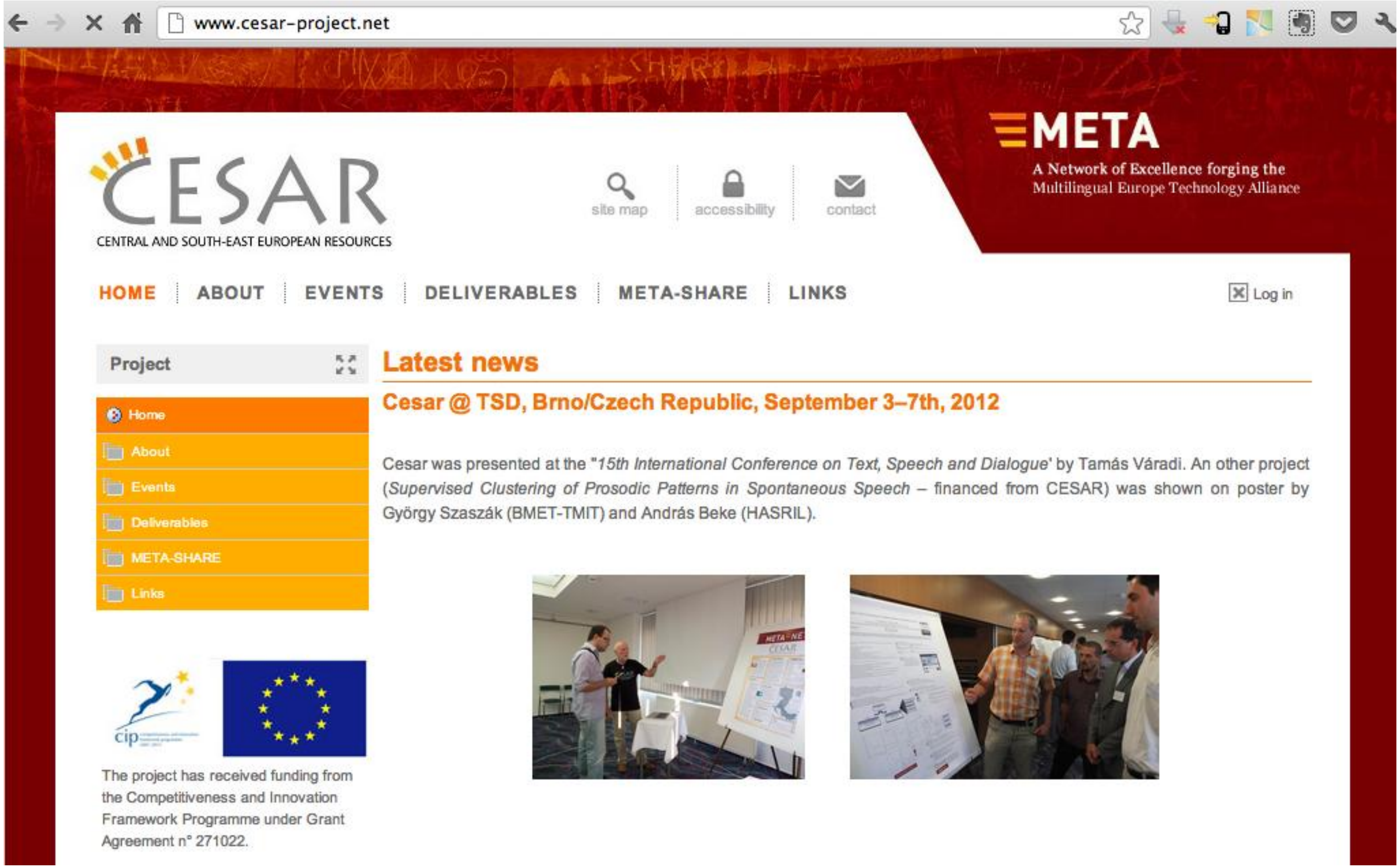
- HLT Days 2012 Sep 07, 2012
- META-FORUM 2012 – Report and Videos now online Sep 04, 2012
- CHAT 2012: Call for Participation May 31, 2012
- More...

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

CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES



Latest news

Cesar @ TSD, Brno/Czech Republic, September 3–7th, 2012

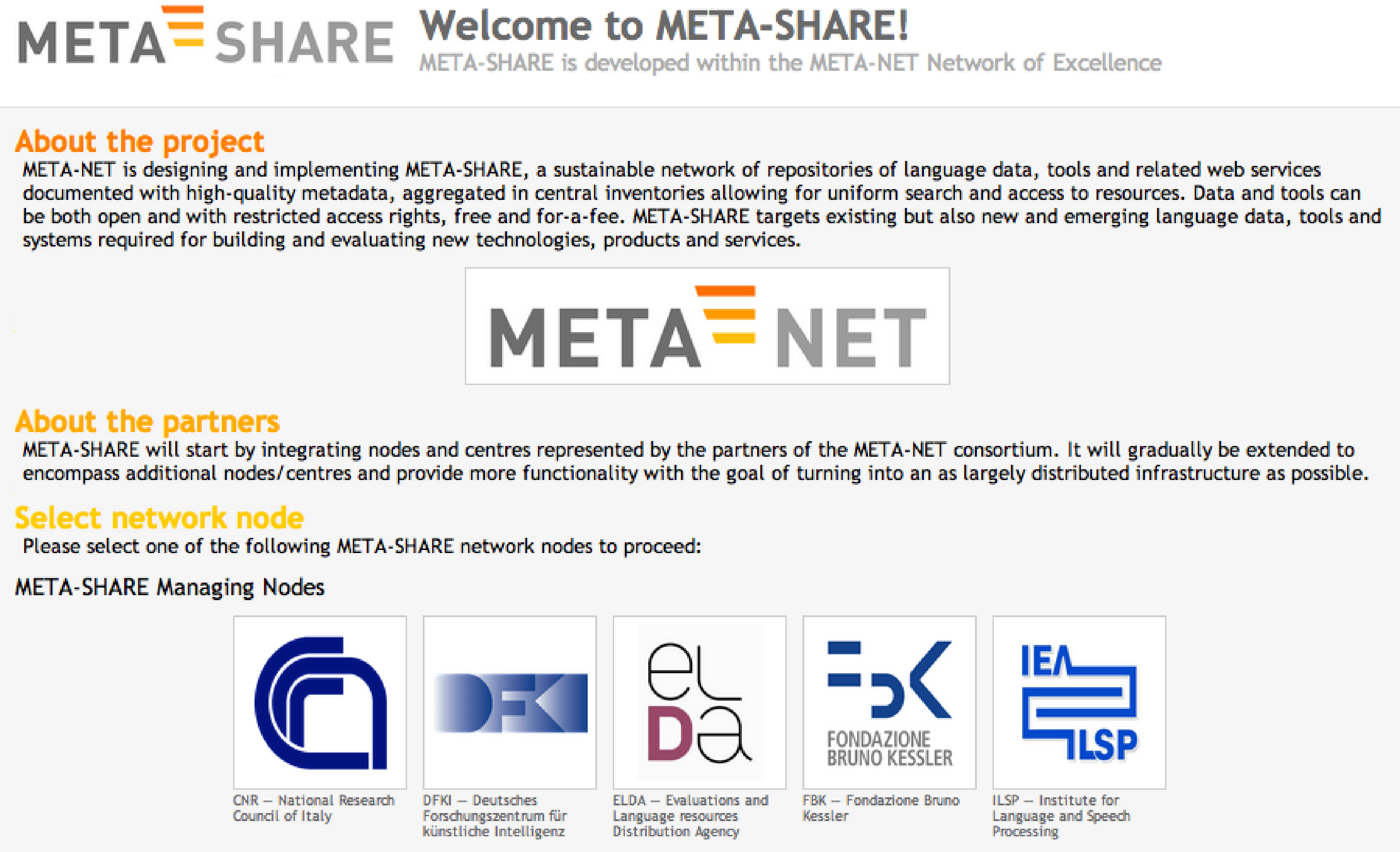
Cesar was presented at the "*15th International Conference on Text, Speech and Dialogue*' by Tamás Váradi. An other project (*Supervised Clustering of Prosodic Patterns in Spontaneous Speech* – financed from CESAR) was shown on poster by György Szaszák (BMET-TMIT) and András Beke (HASRIL).

The project has received funding from the Competitiveness and Innovation Framework Programme under Grant Agreement n° 271022.

CESAR in META-SHARE

www.meta-net.org

CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES



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

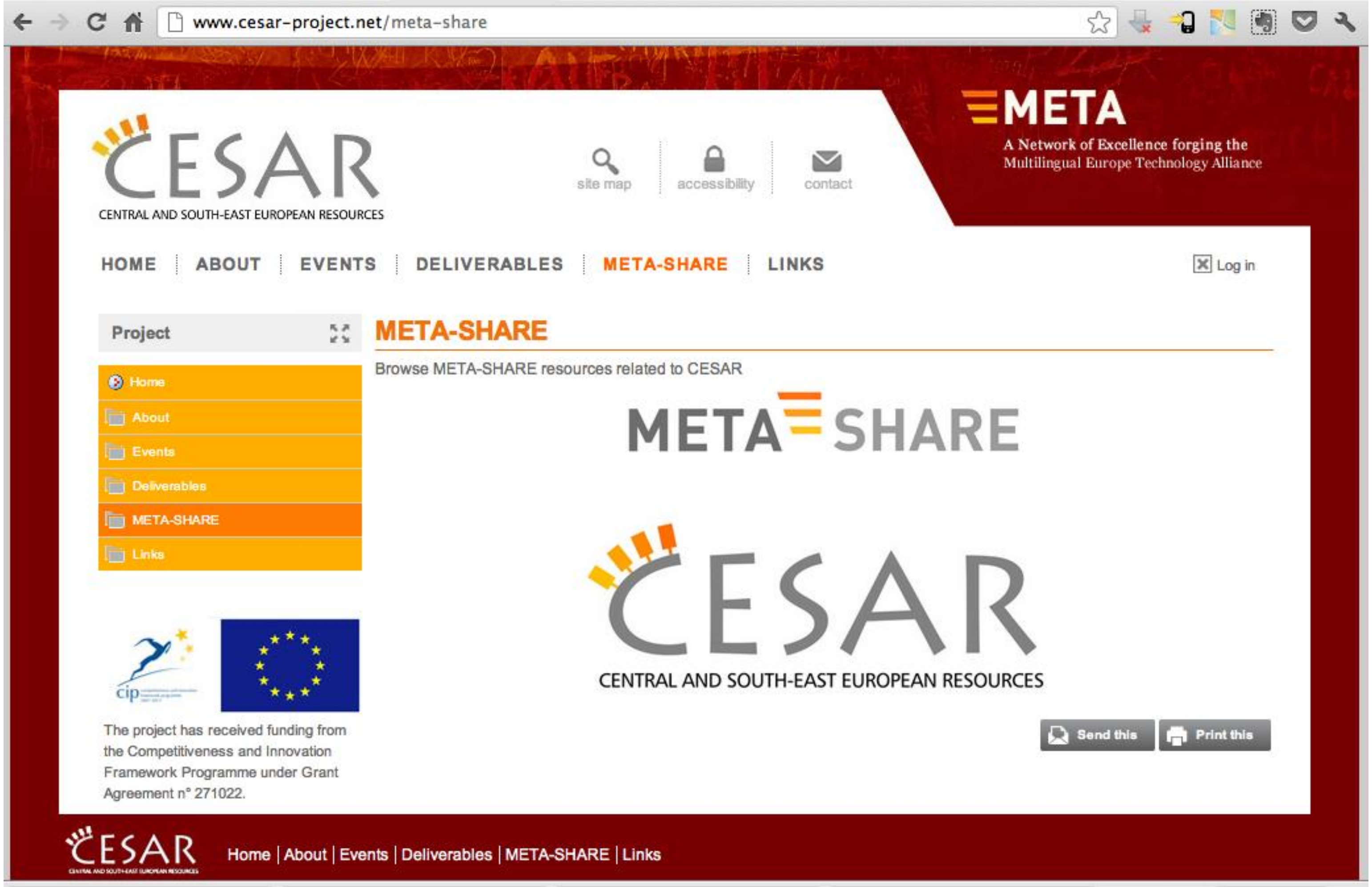
DFKI – Deutsches Forschungszentrum für künstliche Intelligenz

ELDA – Evaluations and Language resources Distribution Agency

FBK – Fondazione Bruno Kessler

ILSP – Institute for Language and Speech Processing

www.cesar-project.net/metashare



www.cesar-project.net/meta-share

CESAR CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES

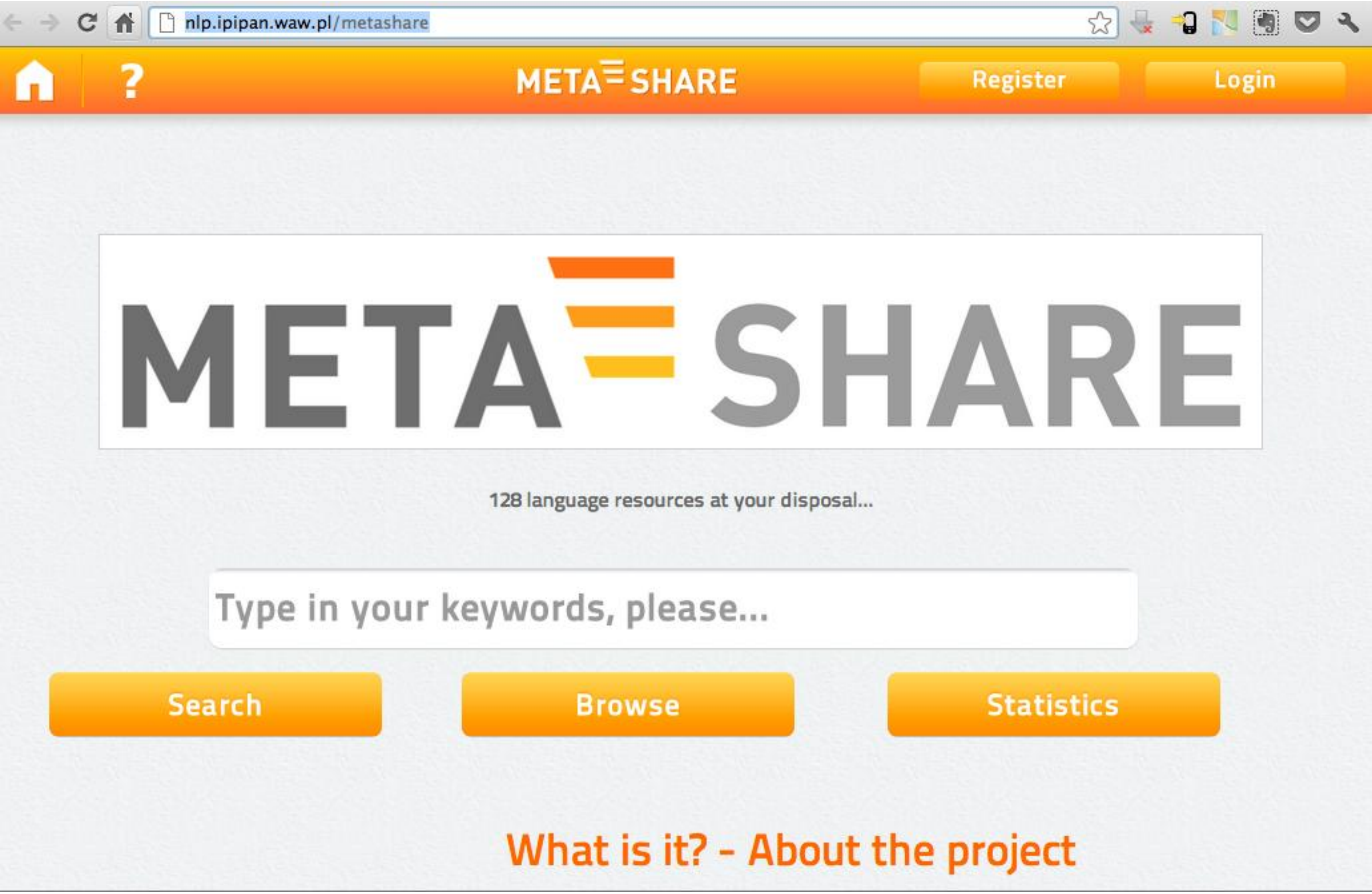
META-SHARE

Browse META-SHARE resources related to CESAR

META-SHARE

CESAR CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES

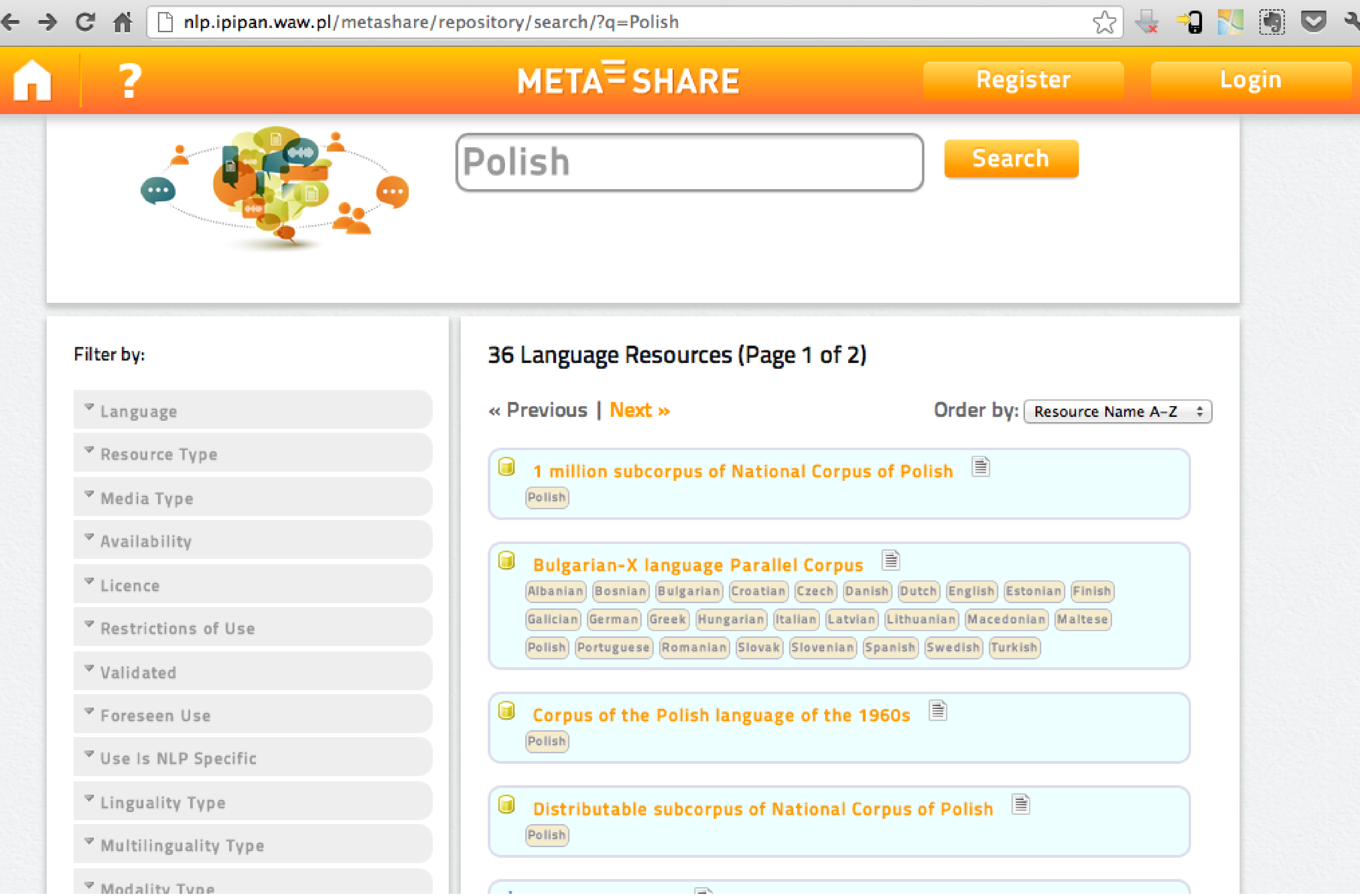
The project has received funding from the Competitiveness and Innovation Framework Programme under Grant Agreement n° 271022.



nlp.ipipan.waw.pl/metashare

128 language resources at your disposal...

What is it? - About the project



Filter by:

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

36 Language Resources (Page 1 of 2)

« Previous | Next »

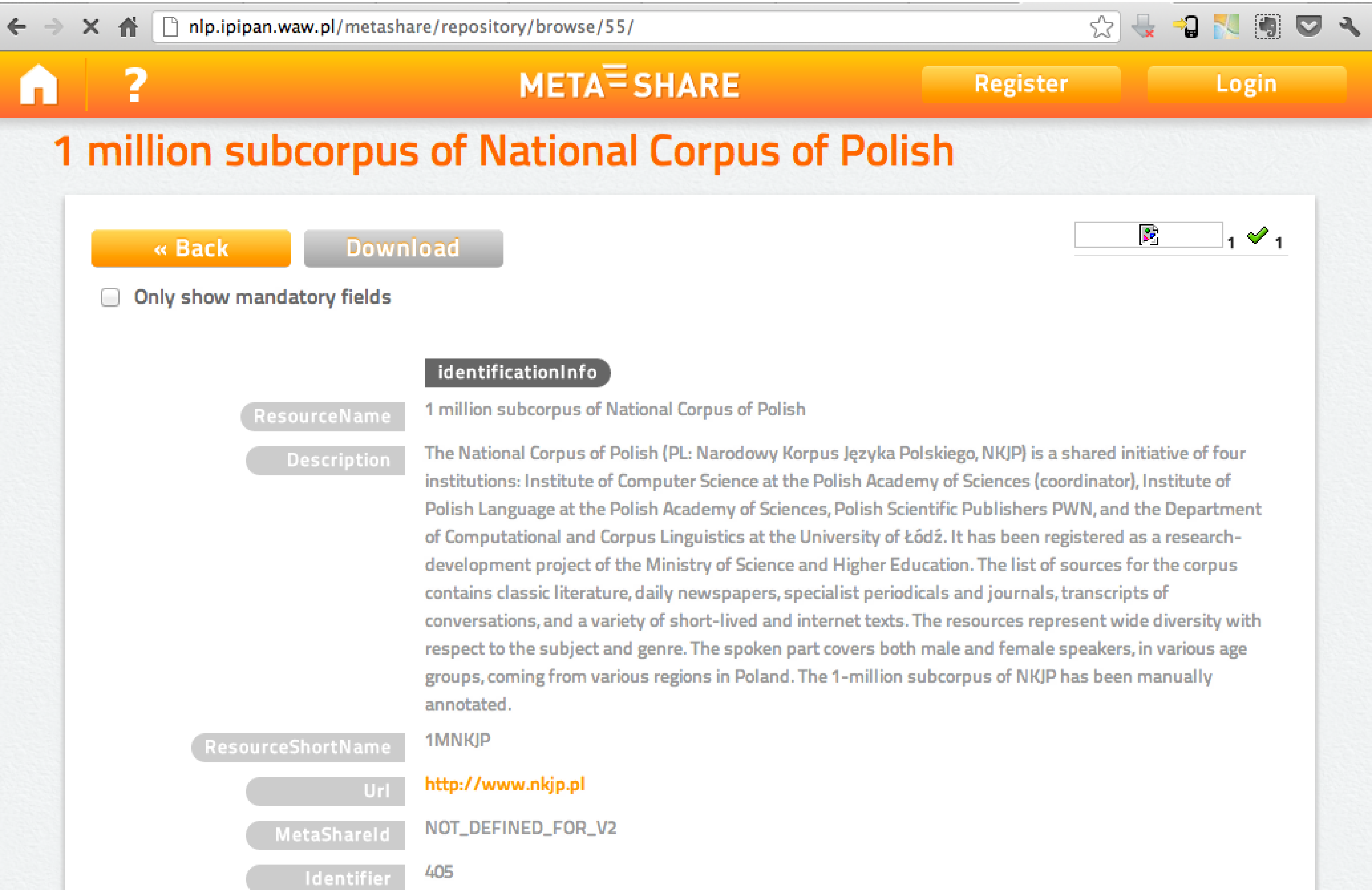
Order by: Resource Name A-Z

1 million subcorpus of National Corpus of Polish
Polish

Bulgarian-X language Parallel Corpus
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

Corpus of the Polish language of the 1960s
Polish

Distributable subcorpus of National Corpus of Polish
Polish



1 million subcorpus of National Corpus of Polish

□ Only show mandatory fields

identificationInfo

ResourceName: 1 million subcorpus of National Corpus of Polish

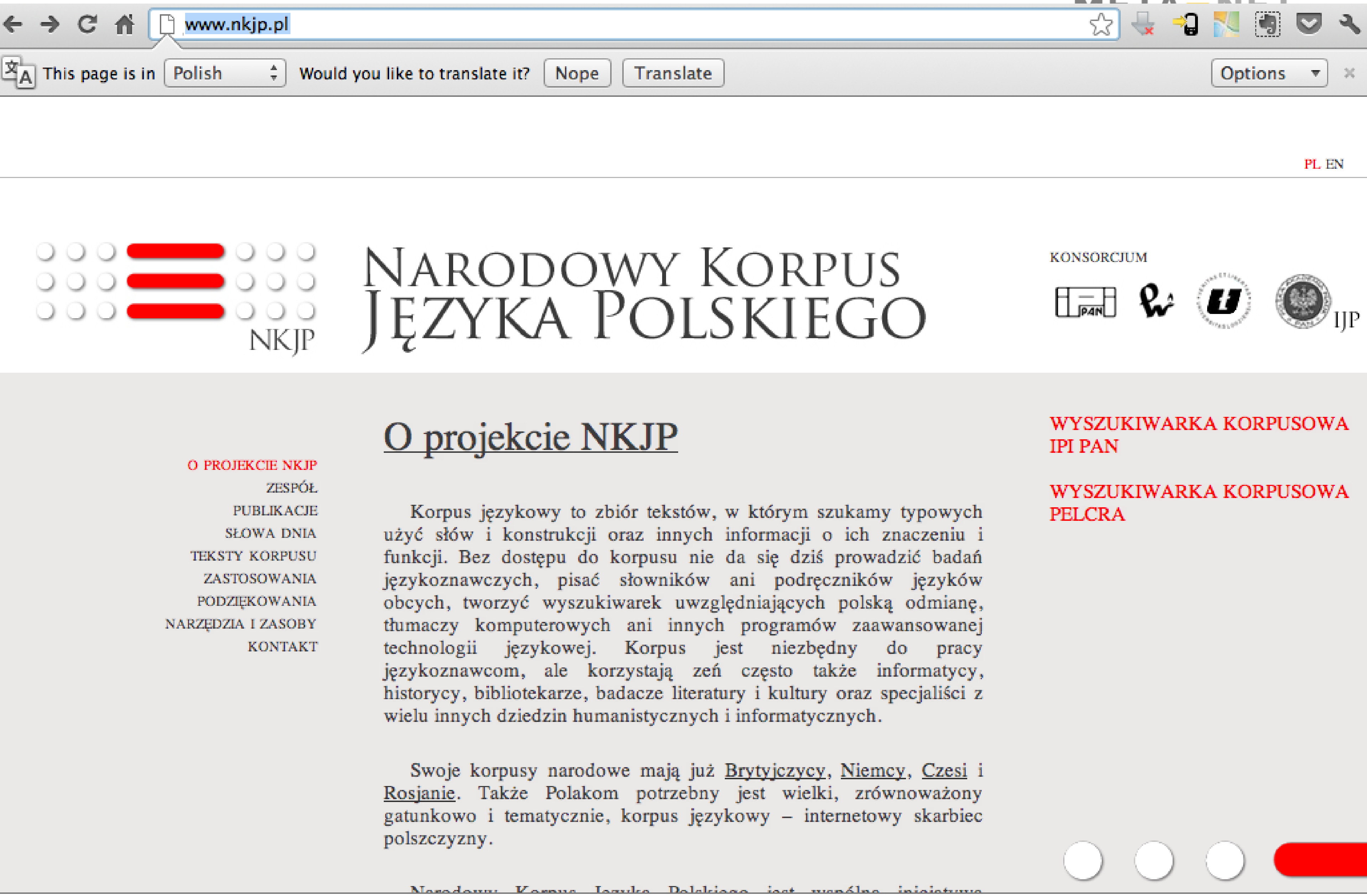
Description: The National Corpus of Polish (PL: Narodowy Korpus Języka Polskiego, NKJP) is a shared initiative of four institutions: Institute of Computer Science at the Polish Academy of Sciences (coordinator), Institute of Polish Language at the Polish Academy of Sciences, Polish Scientific Publishers PWN, and the Department of Computational and Corpus Linguistics at the University of Łódź. It has been registered as a research-development project of the Ministry of Science and Higher Education. The list of sources for the corpus contains classic literature, daily newspapers, specialist periodicals and journals, transcripts of conversations, and a variety of short-lived and internet texts. The resources represent wide diversity with respect to the subject and genre. The spoken part covers both male and female speakers, in various age groups, coming from various regions in Poland. The 1-million subcorpus of NKJP has been manually annotated.

ResourceShortName: 1MNKJP

Url: http://www.nkjp.pl

MetaShareId: NOT_DEFINED_FOR_V2

Identifier: 405



PL EN

NKJP

Narodowy Korpus Języka Polskiego

KONSORCJUM

IJP

- O PROJEKCIE NKJP
- ZESPÓŁ
- PUBLIKACJE
- SŁOWA DNIA
- TEKSTY KORPUSU
- ZASTOSOWANIA
- PODZIĘKOWANIA
- NARZĘDZIA I ZASOBY
- KONTAKT

O projekcie NKJP

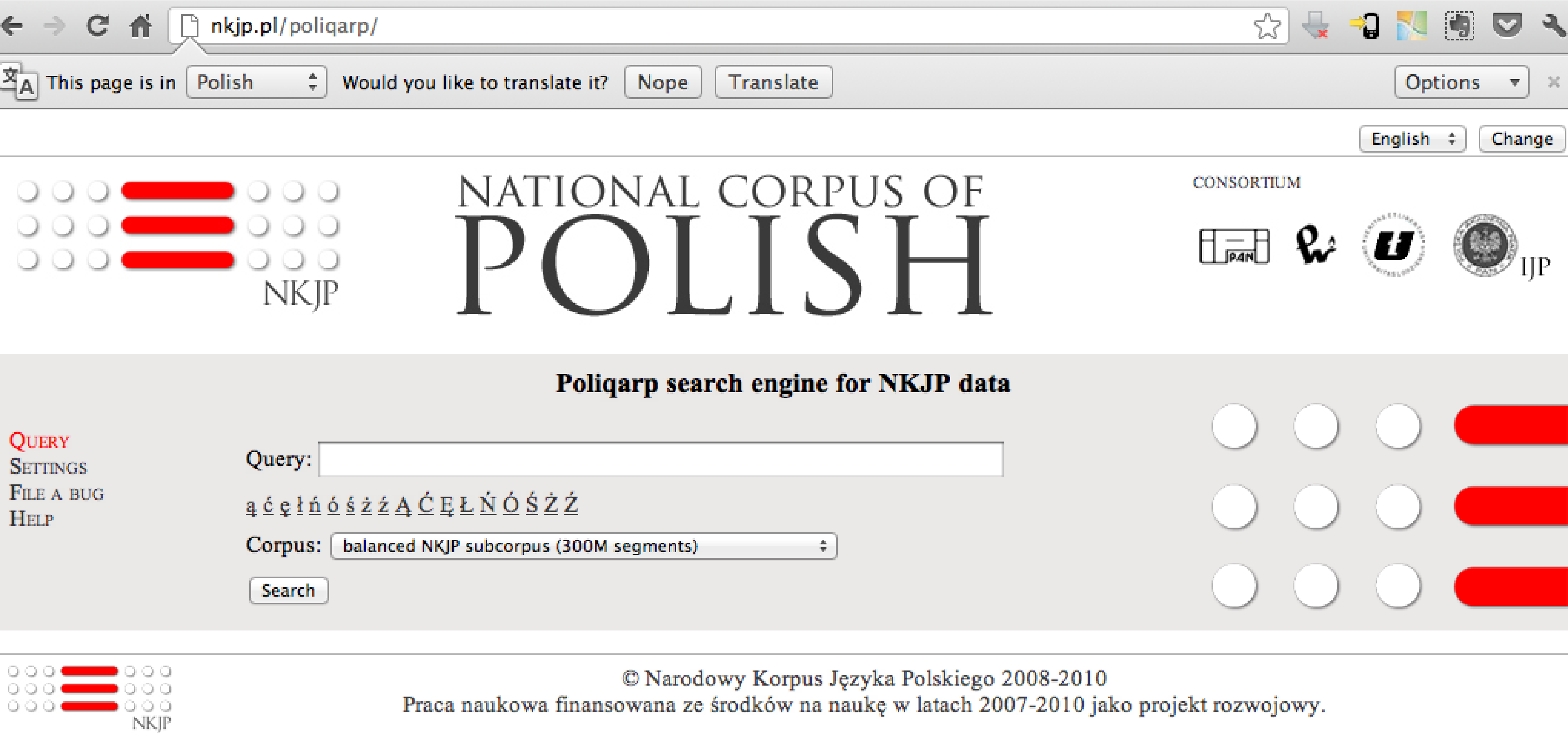
Korpus językowy to zbiór tekstów, w którym szukamy typowych użyć słów i konstrukcji oraz innych informacji o ich znaczeniu i funkcji. Bez dostępu do korpusu nie da się dziś prowadzić badań językoznawczych, pisać słowników ani podręczników języków obcych, tworzyć wyszukiwarek uwzględniających polską odmianę, tłumaczy komputerowych ani innych programów zaawansowanej technologii językowej. Korpus jest niezbędny do pracy językoznawcom, ale korzystają zeń często także informatycy, historycy, bibliotekarze, badacze literatury i kultury oraz specjaliści z wielu innych dziedzin humanistycznych i informatycznych.

Swoje korpusy narodowe mają już Brytyjczycy, Niemcy, Czesi i Rosjanie. Także Polakom potrzebny jest wielki, zrównoważony gatunkowo i tematycznie, korpus językowy – internetowy skarbiec polszczyzny.

Narodowy Korpus Języka Polskiego jest wspólną inicjatywą

WYSZUKIWARKA KORPUSOWA IPI PAN

WYSZUKIWARKA KORPUSOWA PELCRA



NATIONAL CORPUS OF POLISH

Poliqarp search engine for NKJP data

Query:

ą ć ę ł ń ó ś ż ź Ą Ć Ę Ł Ń Ó Ś Ż Ź

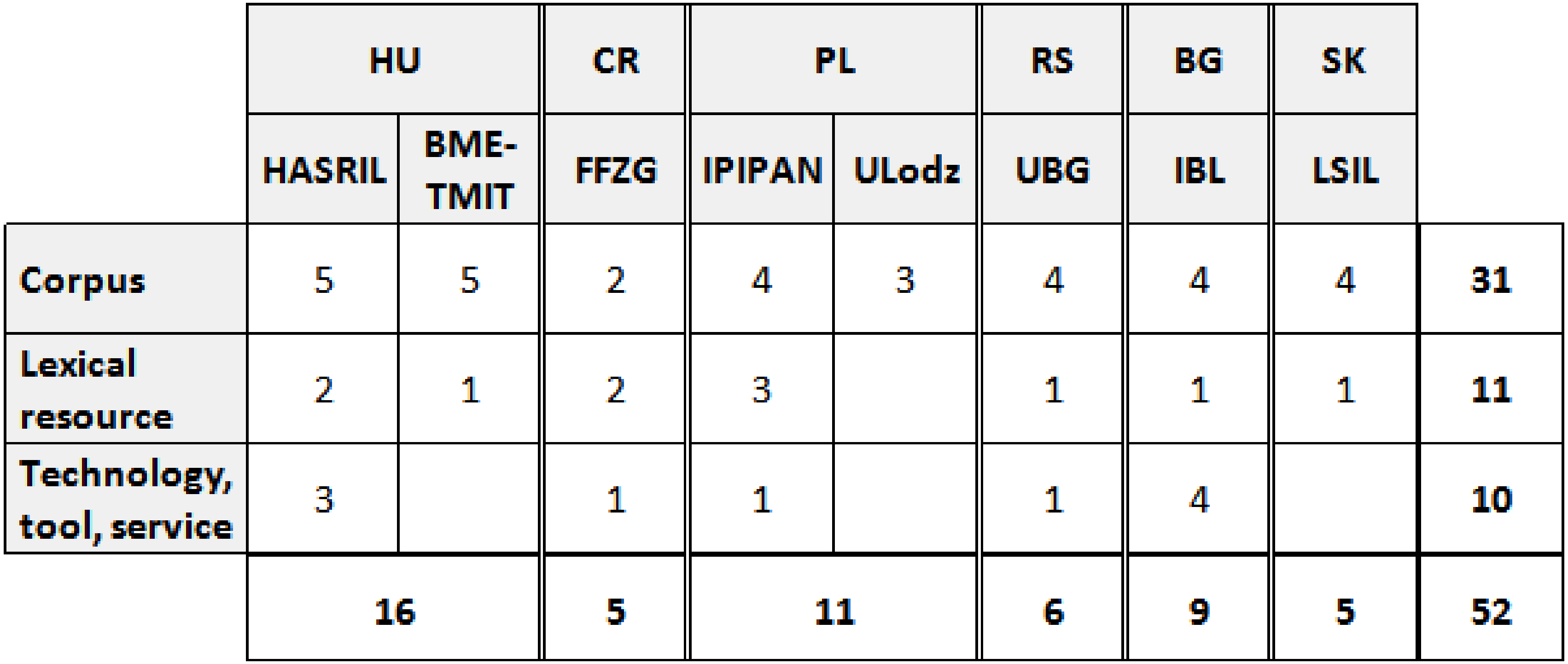
Corpus: balanced NKJP subcorpus (300M segments)

© Narodowy Korpus Języka Polskiego 2008-2010
Praca naukowa finansowana ze środków na naukę w latach 2007-2010 jako projekt rozwojowy.

Results – M18

CESAR First Batch of Resources

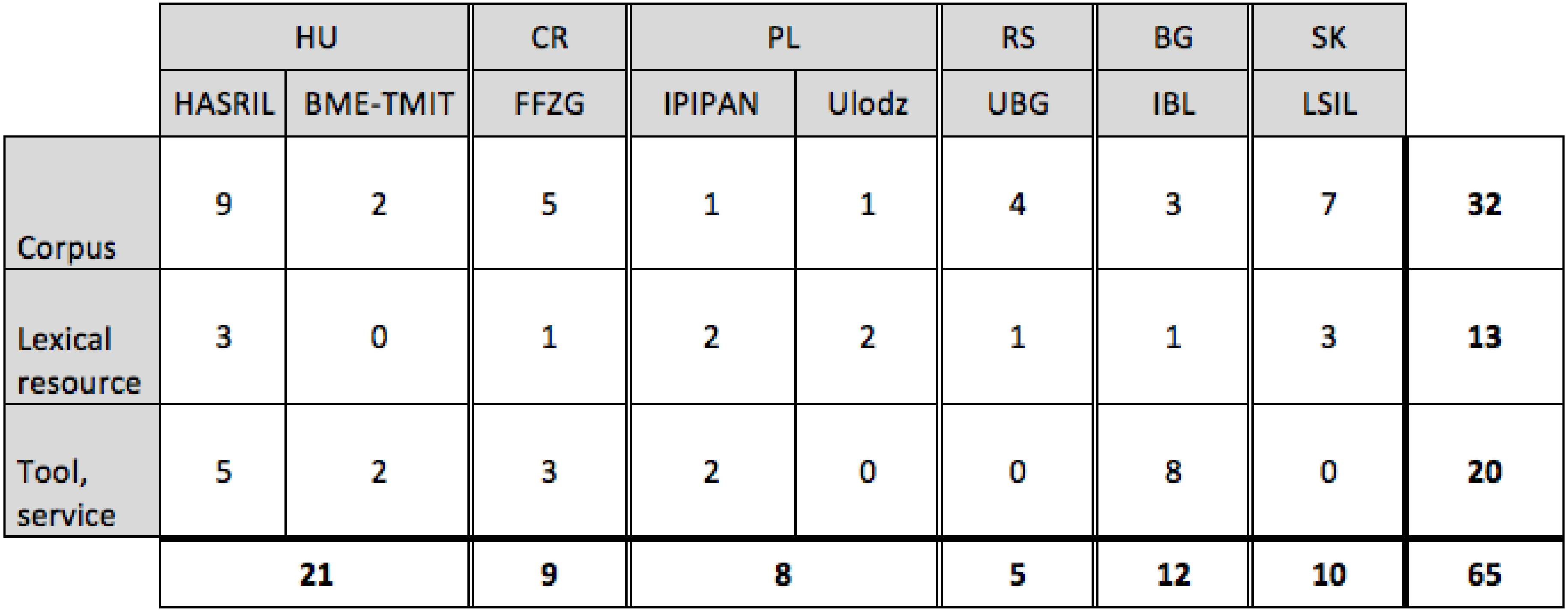
Statistics of resources:



	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

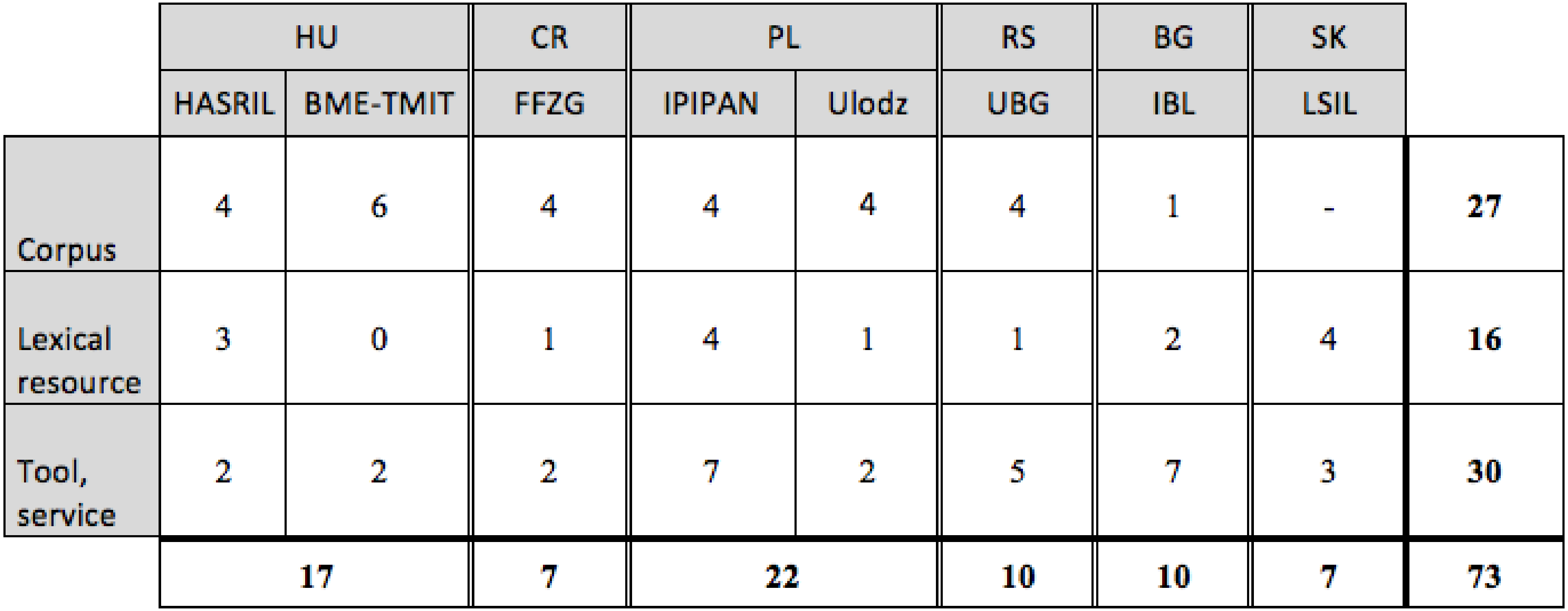
Statistics of resources:



	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**

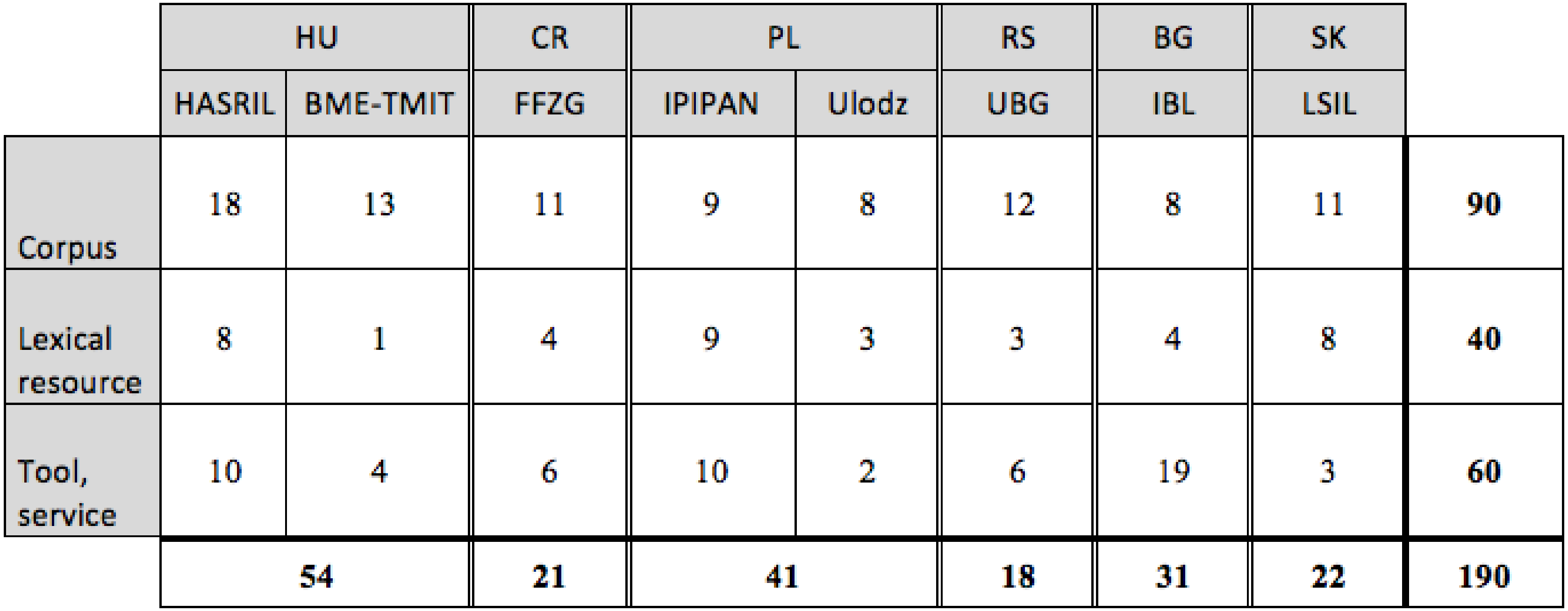
CESAR Third Batch of Resources

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

Polish resources in the 1st Batch

- **IPIPAN**
  - Polish Sejm Corpus
  - PoliMorf Inflectional Dictionary
  - Polish WordNet
  - Polish Named Entity Recognition Tool
  - 1 million subcorpus of National Corpus of Polish
  - Polish Named Entity Resources
  - LUNA.PL Corpus
  - LUNA-WOZ.PL Corpus
- **ULodz**
  - PELCRA Polish-English parallel corpora
  - PELCRA Polish-English parallel corpora
  - PELCRA Polish spoken corpus

Polish resources in the 2nd Batch

- **IPIPAN**
  - Polish Sejm Corpus
  - PoliMorf Inflectional Dictionary
  - Polish WordNet
  - Polish Named Entity Recognition Tool
  - 1 million subcorpus of National Corpus of Polish
  - Polish Named Entity Gazetteer
  - LUNA.PL Corpus
  - LUNA-WOZ.PL 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
  - 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
  - morfologik-stemming
  - Corpus of the Polish language of the 1960s
  - Shallow Grammar for the National Corpus of Polish
  - PANTERA
  - PolNet

Polish resources in the 2nd Batch

- **Ulodz**
  - PELCRA Polish-English parallel corpora (CC-BY)
  - PELCRA Polish-English parallel corpora (CC-BY-NC)
  - PELCRA Polish spoken corpus (CC-BY-NC)
  - ECL Dictionaries
  - PELCRA EN Lemmatizer
  - PELCRA Language Detector
  - PELCRA Polish-English parallel corpus of literary works (CC-BY)
  - PELCRA mutlilingual 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

Proposed Polish resources in the 3rd Batch

- Składnica
- The Corpus of Polish Summaries
- The Parallel English-Polish Corpus
- Redistributable Polish-Russian Corpus
- Learner Speech Database
- SNUV voice recognition speech database
- PELCRA Time-Aligned Spoken Corpus
- Paralela DB

- Polish Open CYC lexicon
- Polish-English Wikipedia NE dictionaries
- Lexeme Forge
- Slowal
- Lakon
- Świgra
- Ruler
- PolSumm
- VOICE LAB Automated Speech Recognition (ASR) engine

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

Gaps and Challenges*

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

META-NET Language Whitepapers

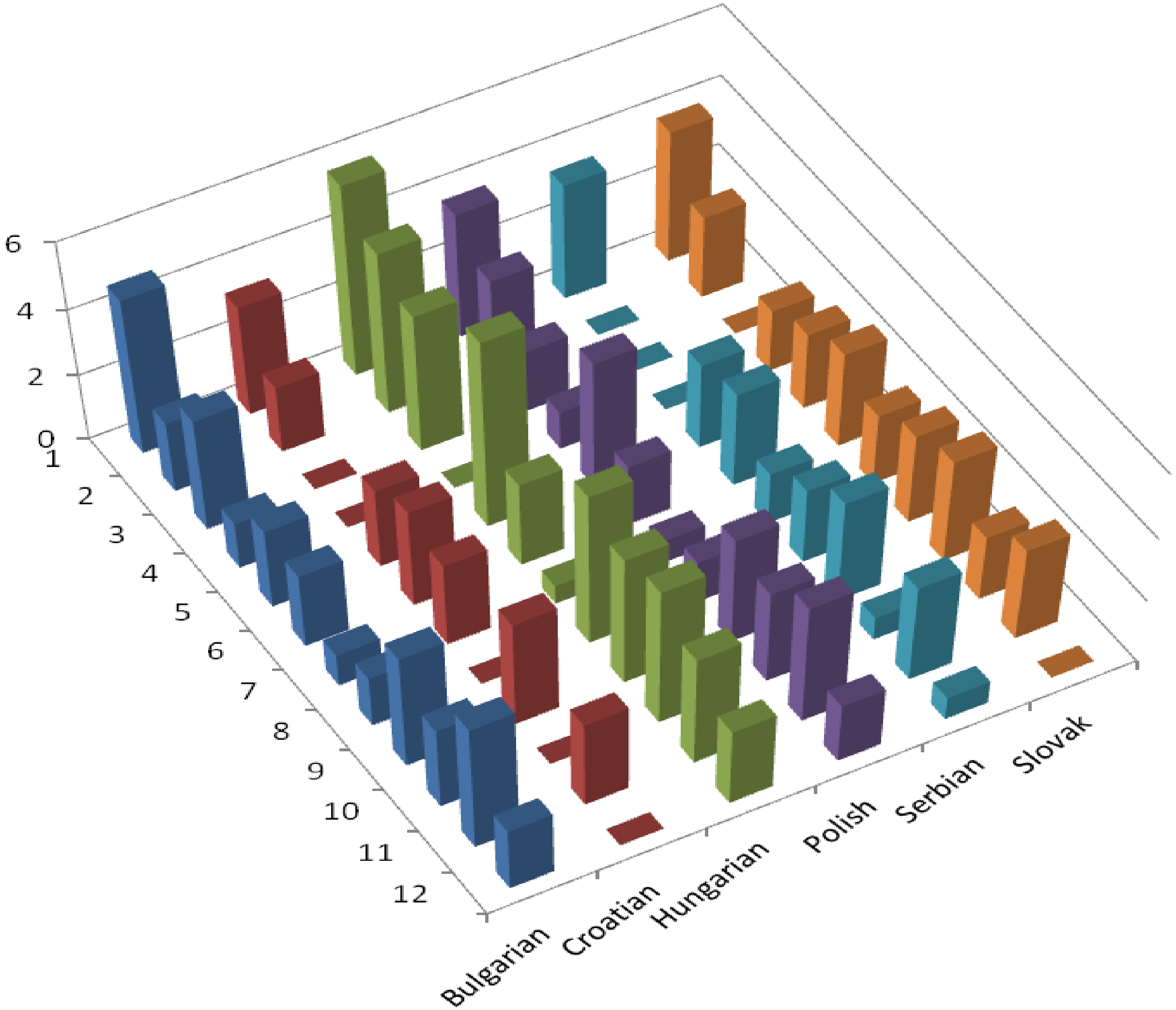
- 1st edition
  - 30 European languages
  - META-FORUM, Budapest, 2011-06
- two tables from 1st edition
  - resources
  - tools/services
- tables with non-merged categories used
  - more detailed list of LR&T categories present
  - allows for more finegrained detection
- used as a data source about CESAR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

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
- 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

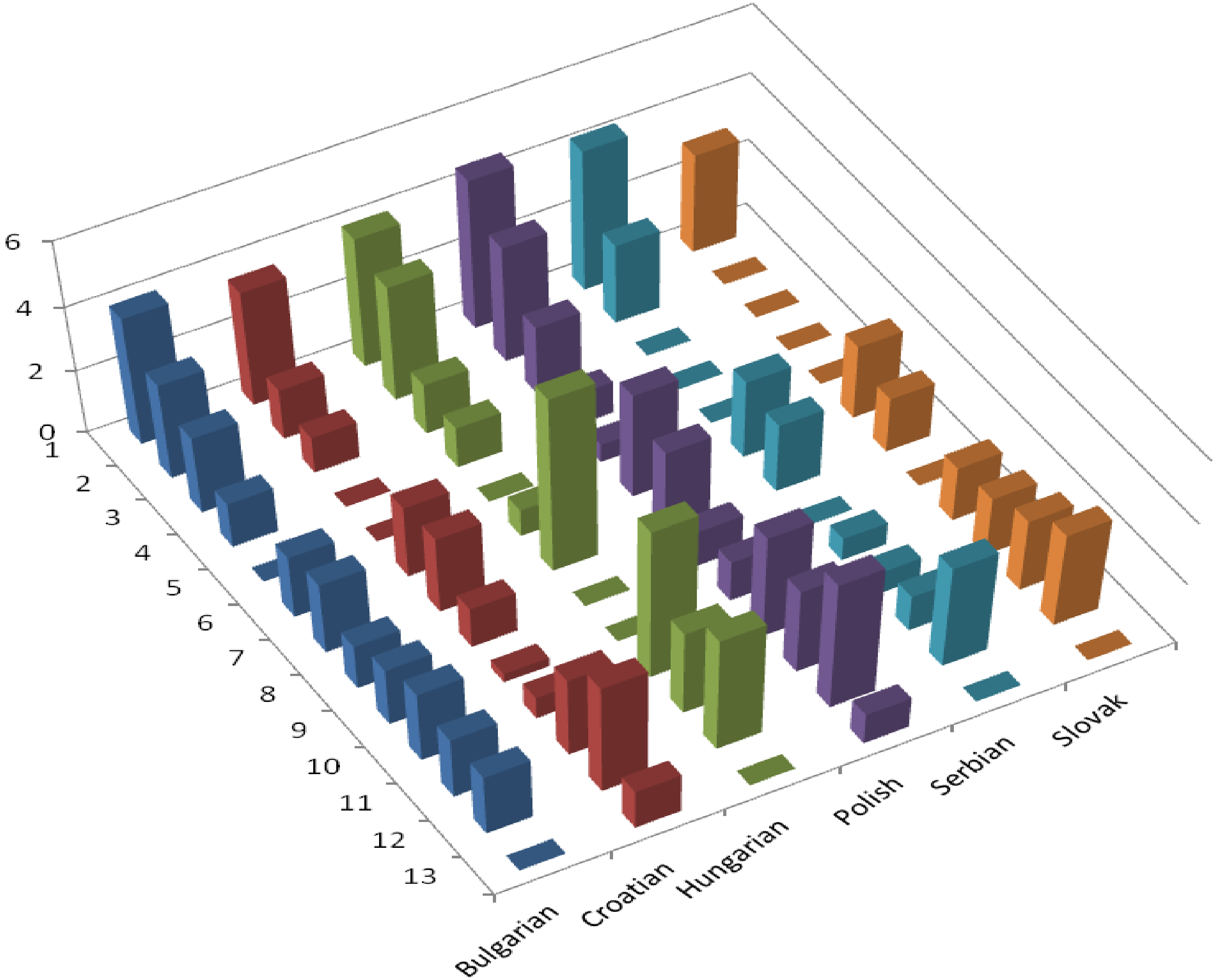
  - category well defined?
  - misunderstanding in criteria for giving scores between national experts?
- similar discrepancy does not appear in other categories
- individual languages snapshot
  - vertical reading of the table

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
- Polish 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