CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES

The Contribution of CESAR to META-SHARE

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

SEVENTH FRAMEWORK PROGRAMME

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

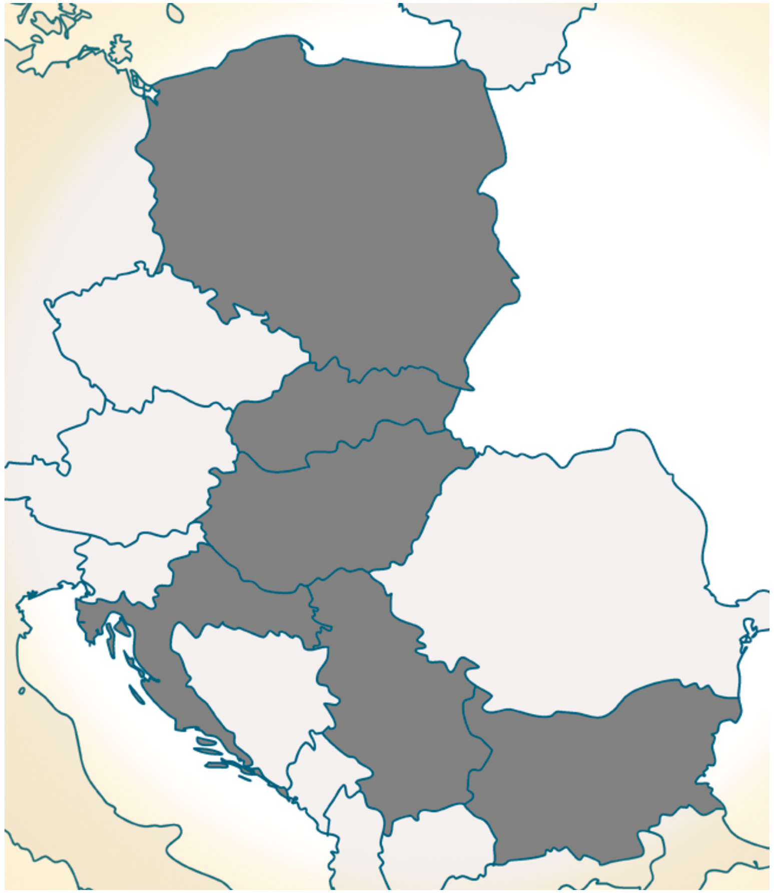
- the CESAR project in a nutshell
- CESAR in META-SHARE
- First Batch
- Second Batch
- Third Batch
- Beyond

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

CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES

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
- **bridge the technological gap between this region and the other parts of Europe**
  - **filling obvious and important blind spots in language resources and tools infrastructure**

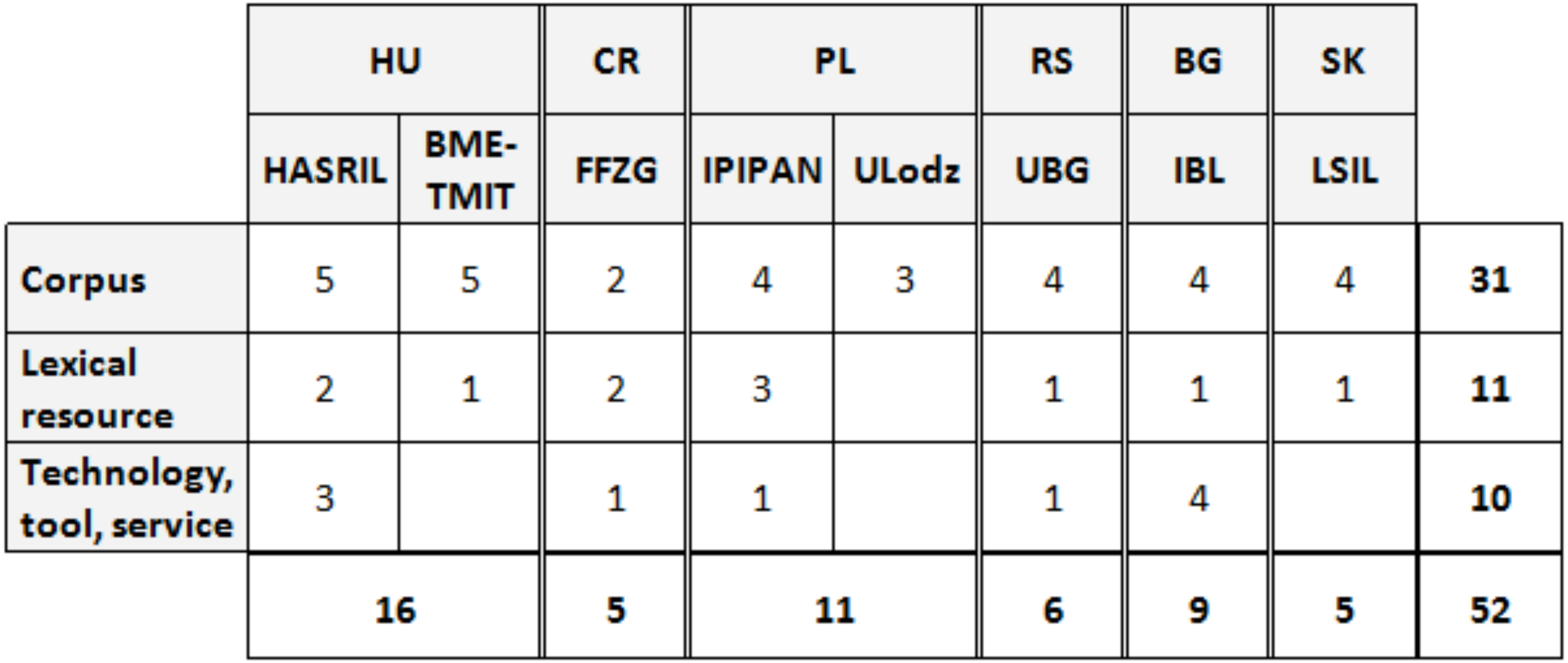
Timeline

- 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**

First-year results

CESAR First Batch of Resources

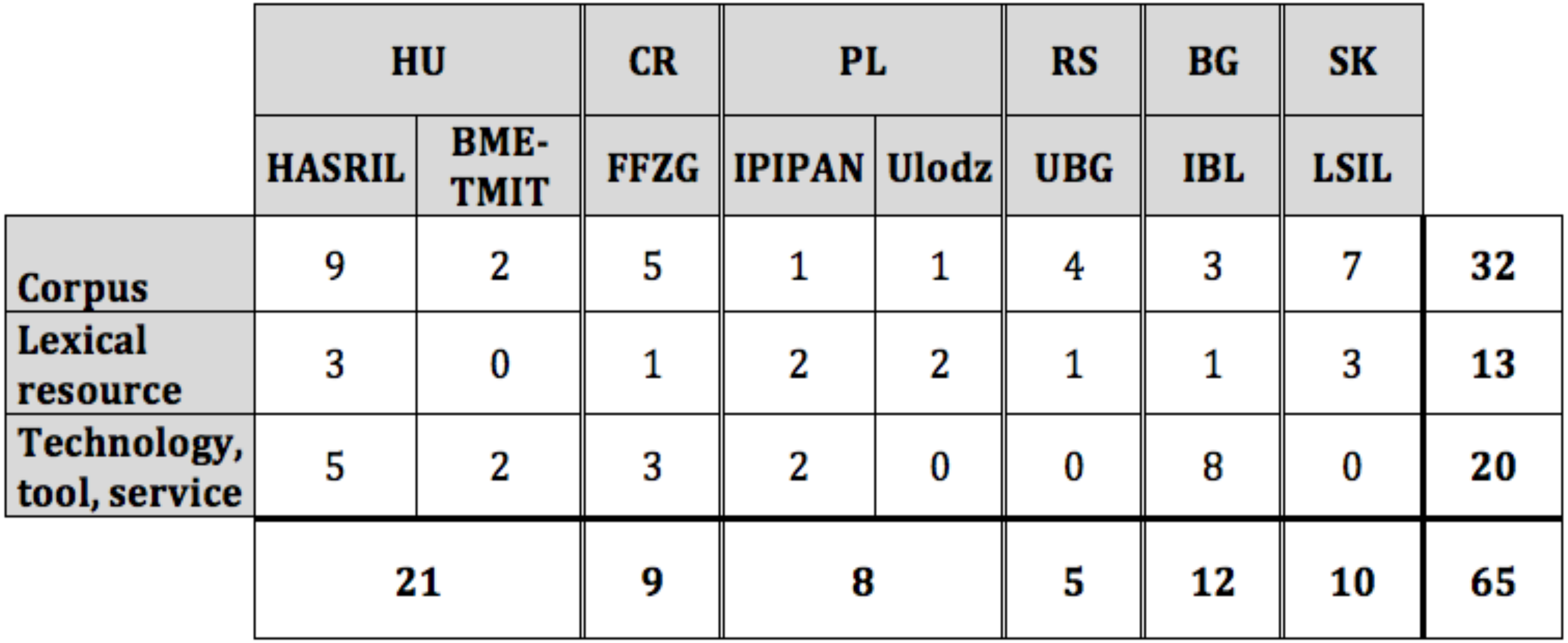
Statistics of resources:



	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**

CESAR Second Batch of Resources

Statistics of resources:



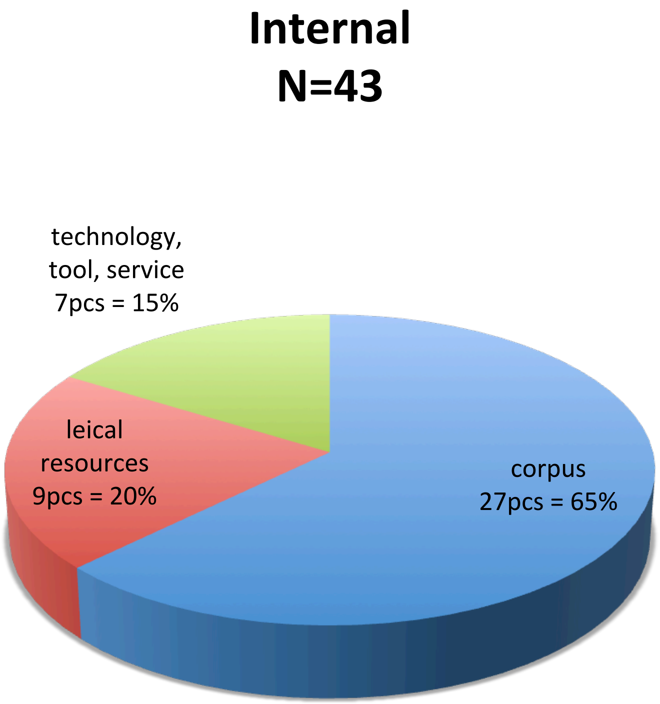
	HU		CR	PL		RS	BG	SK	
	HASRIL	BME-TMIT	FFZG	IPIPAN	Ulodz	UBG	IBL	LSIL	
Corpus	9	2	5	1	1	4	3	7	32
Lexical resource	3	0	1	2	2	1	1	3	13
Technology, tool, service	5	2	3	2	0	0	8	0	20
	21		9	8		5	12	10	65

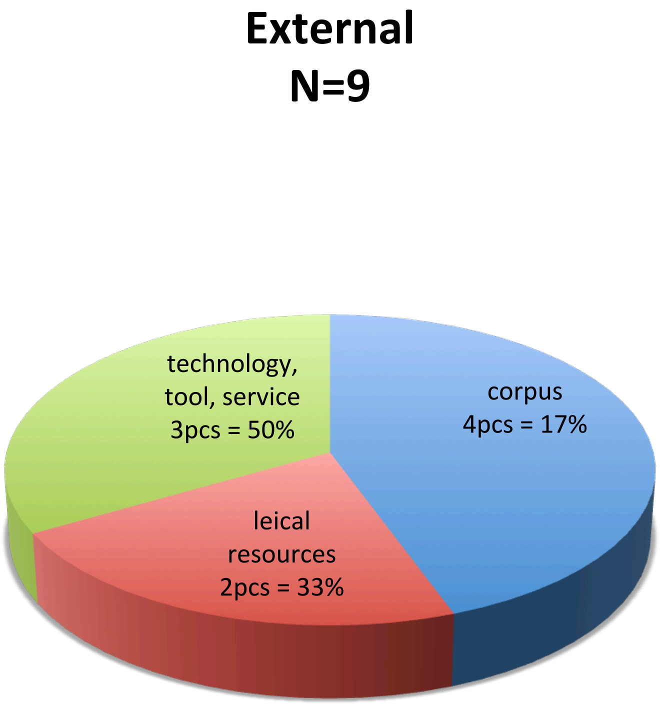
‘In other words’

	monolingual corpus (token)	paralel corpus (token)	record/entry/lexicon
Batch 1	4 335 446 886	127 710 000	3 763 121
Batch 2	1 702 565 806	41 810 000	1 640 579
Total	**6 038 012 692**	**169 520 000**	**5 403 700**

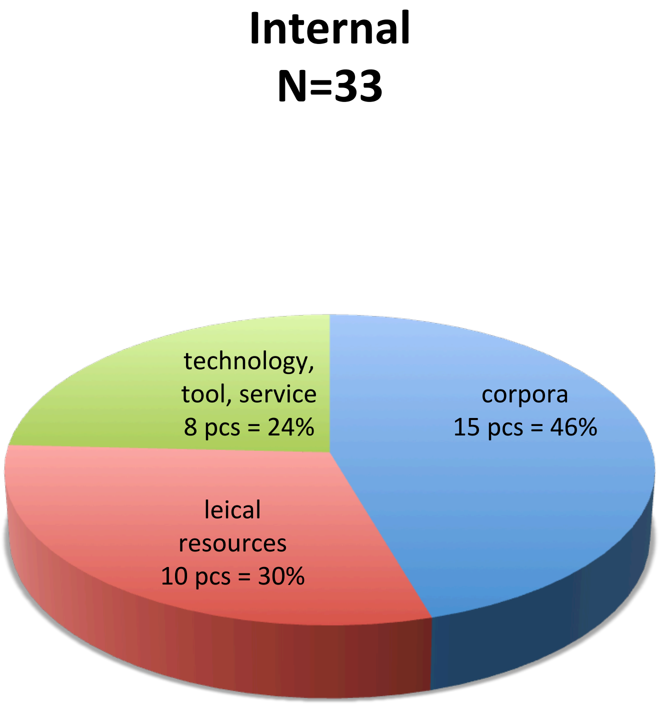
	Corpus	Lexical resource	Technology/ tool/service	**TOTAL**
Batch 1	31	11	10	**52**
Batch 2	32	13	20	**65**
Total	**63**	**24**	**30**	**127**

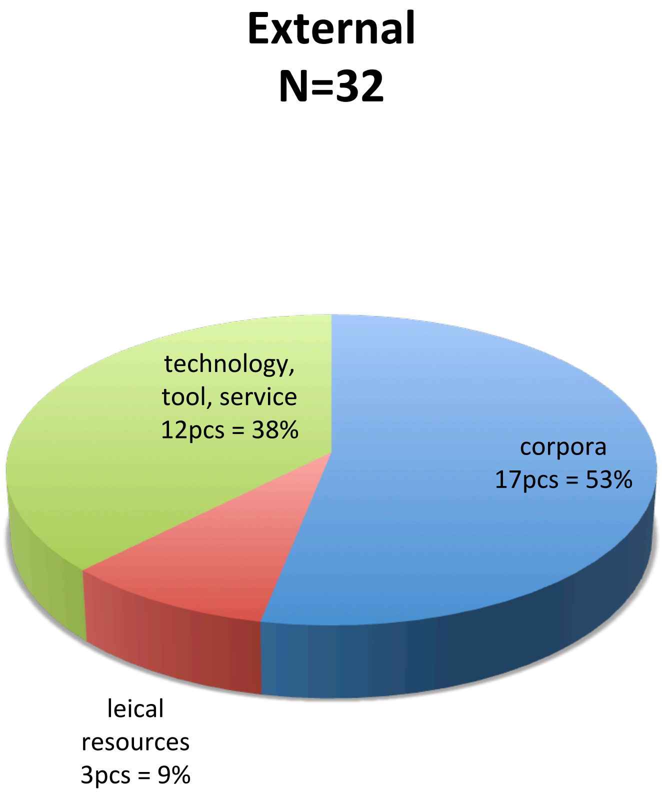
First batch


Internal
N=43
technology,
tool, service
7pcs = 15%
leical
resources
9pcs = 20%
corpus
27pcs = 65%


External
N=9
technology,
tool, service
3pcs = 50%
corpus
4pcs = 17%
leical
resources
2pcs = 33%

Second batch





BIG DATA in CESAR

- Monolingual Corpora
  - Balanced Slovak Corpus - 247 000 000 tokens
  - Slovak National Corpus - 770 000 000 tokens
  - Slovak Legal Texts Corpus - 146 000 000 tokens
- Parallel Corpora
  - Bulgarian X language parallel corpora - 100 000 000 tokens
  - Hunglish (Hungarian English) Corpus - 4 151 000 sentences
  - Polish-Russian Parallel Corpus - 25 000 000 words
- Speech Corpora
  - Mindentudás Speech Corpus (Hungarian) - 200 hours
  - Corpus of Spoken Slovak - 178 hours
  - Broadcast Lectures Database - 150 hours
  - Hungarian Parliamentary Speeches - 1000 hours

Third batch

- **Cross-linked resources for all six languages**
- n-grams derived from national corpora
- Collocational lexica derived from corpora
- multilingual parallel corpora
- multilingual aligned named entity annotated corpus

Long-term perspectives

- **Strong commitment to support META-SHARE**
- In a matter of days **six** of the **nine** partners presented signed **letters of intent**
  - ❑ Set up a META-SHARE repository to host and make available its language resources through the META-SHARE network
  - ❑ continue to host the repository of LRs and serve as a META-SHARE node for 24 months
  - ❑ continue to provide technical and/or user support services for 24 months
  - ❑ start or continue to participate in the META-SHARE software development team

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