# Supporting Less-resourced Languages in Multilingual Europe


## Tamás Váradi

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


Multilingual Digital Content Workshop
Unity in Diversity Conference
Vilnius, 26-09-2013


Co-funded by the 7th Framework Programme and the ICT Policy Support Programme of the European Commission through the contracts T4ME, CESAR, METANET4U, META-NORD (grant agreements no. 249119, 271022, 270893, 270899).

# Outline

CESAR
CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES

- **Challenges**
- **Threats**
- **Achievements**
  - META-NET CESAR Project
  - EFNILEX P

# Less-resourced languages

CESAR
CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES

- Resources: datasets, analysis, software tools
- No standard definition
- Basic Language Resource Kit
- At stake is digital divide or digital extinction
- Speakers of less-resourced languages should enjoy the benefits of digital services in equal manner

# Position of EU Languages

CESAR
CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES

www.meta-net.eu/whitepapers/overview

## META-NET White Paper Series

- **Overview and List of Volumes**
- Quotes and Testimonials
- Authors and Contributors
- Key Results
- Press Release
- Press Coverage
- The Team Behind the Series

### Europe's Languages in the Digital Age

**31 Volumes cover 30 European Languages**

Basque, Bulgarian, Catalan, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, Galician, German, Greek, Hungarian, Icelandic, Irish, Italian, Latvian, Lithuanian, Maltese, Norwegian (bokmål), Norwegian (nynorsk), Polish, Portuguese, Romanian, Serbian, Slovak, Slovene, Spanish, Swedish.

**At a Glance**

- Key Results and Cross-Language Comparison

**Aims and Scope**

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. The benefits offered by Language Technology differ from language to language. So do the actions that need to be taken within META-NET, depending on the factors such as the complexity of the respective language, the size of its community, and the existence of active research centres in this area.

The META-NET Language White Paper series "Languages in the European Information Society" reports on the state of each European language with respect to Language Technology and explains the most urgent risks and chances. The series will cover all official European languages and several other languages spoken in geographical Europe. While there have been a number of valuable and comprehensive scientific studies on certain aspects of languages and technology, there exists no generally understandable compendium that takes a stand by presenting the main

# Position of CESAR languages

CESAR
CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES

| | | | | | |
|---|---|---|---|---|---|
| **MT** | | 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 |
| **Text Analysis** | 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 |
| **Speech** | 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 |
| **Resources** | 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 |

# Challenges

CESAR
CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES

- Size of population, size of market
- LT capacities are scarce and fragmented
- Skills and training are often lacking
- Off-the-shelf solutions do not work
- Return on investment makes R&D economically not viable

# Threats

CESAR
CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES

- Global players a mixed blessing:
  - Raise false expectations
  - Go for “just good-enough” solutions
  - Impose business models that are not viable
  - Create a false sense of comfort for decision makers

# META-NET CESAR Project

CESAR
CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES

- **CESAR stands for CEntral and Southeast EuropeAn Resources**
- **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)
- **cc. 95 m speakers**
- **all languages Slavic, except Hungarian**

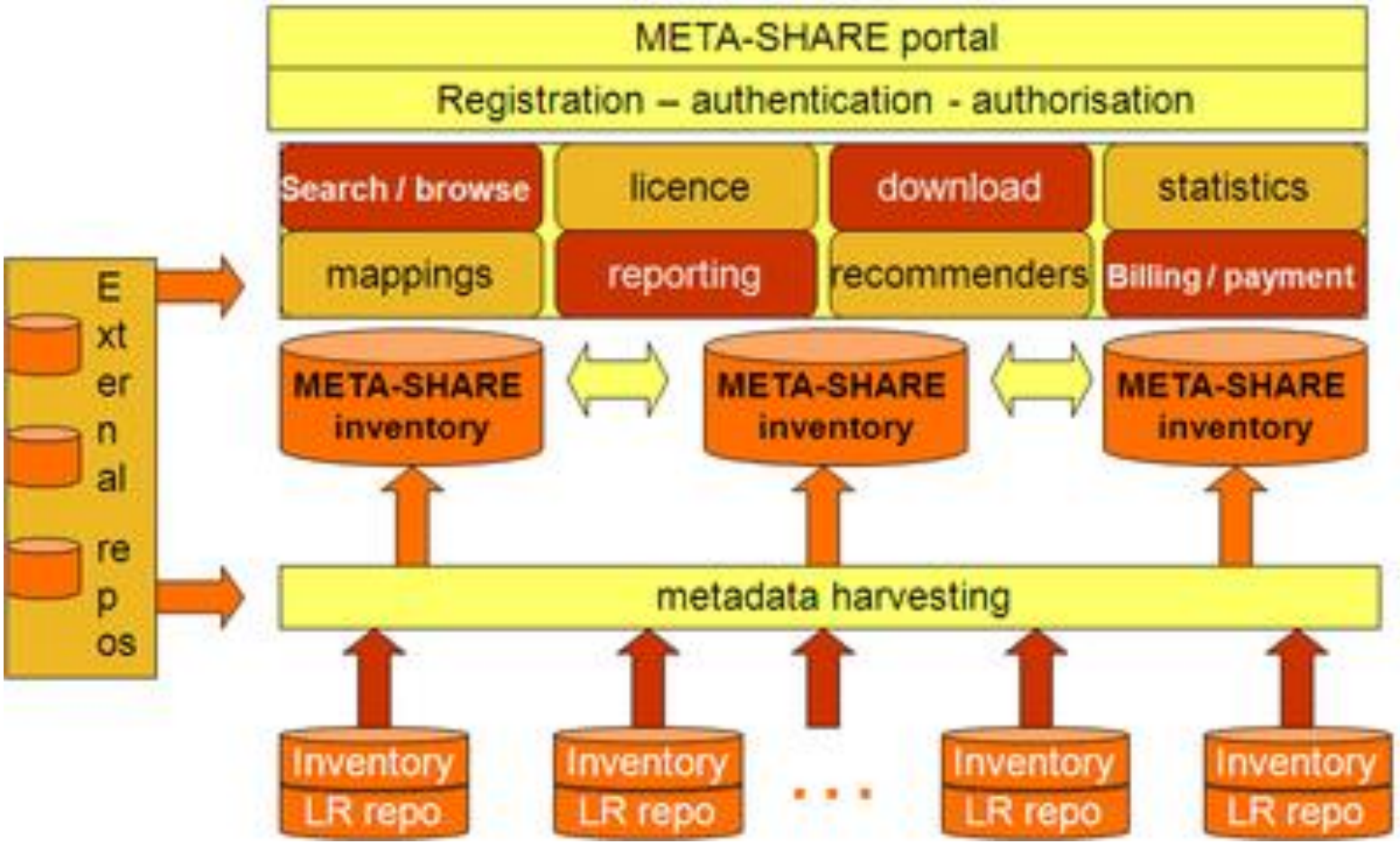
# META-SHARE

CESAR
CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES

- Open
- Integrated
- Distributed
- Language resources and tools infrastructure

# META-SHARE architecture

CESAR
CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES



# Distribution of total resources

CESAR
CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES

| | HU | | CR | PL | | RS | | BG | SK | |
|---|---|---|---|---|---|---|---|---|---|---|
| | HASRIL | BME-TMIT | FFZG | IPIPAN | Ulodz | UBG | IPUP | IBL | LSIL | |
| **Corpus** | 19 | 21 | 12 | 17 | 11 | 10 | | 9 | 21 | **120** |
| **Lexical/Conceptual resource** | 6 | 1 | 9 | 23 | 1 | 3 | 2 | 11 | 9 | **65** |
| **Technology, tool, service** | 6 | 3 | 5 | 19 | 5 | 6 | | 16 | 6 | **66** |
| | **56** | | **26** | **76** | | **21** | | **36** | **36** | **251** |

# EFNILEX

CESAR
CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES

- **Objective:** provide LT support for dictionaries between commercially not viable language pairs
- **Method:** word alignment from parallel corpora (translated texts)
- **Benefits:** lexical equivalents with sense alignment, example sentences from corpora
- **Results:** online, customisable tool for Lithuanian-Hungarian, Dutch-French, English-Hungarian, Slovanian-Hungarian
  Waiting to be scaling up!

**www.efnilex.efnil.org**

# EFNILEX

CESAR
CENTRAL AND SOUTH-EAST EUROPEAN RESOURCES

EFNIL
European Federation of National Institutions for Language
www.efnil.org

RESEARCH INSTITUTE FOR LINGUISTICS
HUNGARIAN ACADEMY OF SCIENCES

Dictionary Browser
Cut Board
Info
Help

English | Hungarian

**Under construction. The page may have bugs.**

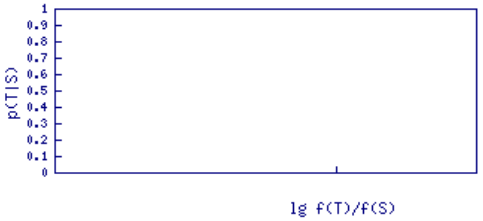
*Should you have any comments or suggestions, please send an email to Enikő Héja at heja.eniko@nytud.mta.hu.*

Language pairs: hu-en | en-hu | hu-sl | sl-hu | hu-lt | lt-hu | fr-nl | nl-fr

Criteria: max f(S) / f(T) = 10, max f(T) / f(S) = 10, min f(S) = min f(T) = 5, min p(T|S) = 0.008,
$f_1(S) \in [1, 10] \rightarrow$ min p(T|S) = 0.3, $f_2(S) \in [10, 100] \rightarrow$ min p(T|S) = 0.14, $f_3(S) \in [100, 1000] \rightarrow$ min p(T|S) = 0.04

ügyes_A

| ~ | prob | freq | compo |
|---|---|---|---|
| clever$_A$ | 0.208 | 152 | |
| skilful$_A$ | 0.072 | 36 | |
| skill$_N$ | 0.067 | 157 | |
| handy$_A$ | 0.046 | 37 | |




clever$_A$ handy$_A$ skilful$_A$ skill$_N$

| hu | en |
|---|---|
| Élővíz öntözte dalolva a fákat , és a romba dőlt oszlopsorok , a vad sziklák , amelyeket ügyes építész készített , egy tóban tükröződtek a szobrokkal együtt . | They were watered by a running brook , and colonnades in ruins , and imitation rocks , arranged by a skilful artist , were reflected in a lake , which also mirrored the statues that stood round it . |
| Pürrhosz odadöfte kardját , fejét elfordítva , és ügyes fogás segítségével a vér patakozva ömlött a szűz ragyogó kebléből , és Polüxéna hátrahanyatló fejjel és a halál iszonyatával a szemében , méltóságteljesen esett össze . | Pyrrhus , turning away his head , plunged his sword into her heart , and by a skilful trick , the blood gushed forth over the dazzling white breast of the virgin , who , with head thrown back , and her eyes swimming in the horrors of death , fell with grace and modesty . |
| A végzet - a legügyesebb színpadi rendező - csak ritkán alkot olyan jeleneteket , fejleszti ki úgy a drámát , hogy legalább egy elfogulatlan szemlélő ne legyen a színpadon . | Destiny , it may be , - the most skilful of stage managers , - seldom chooses to arrange its scenes , and carry forward its drama , without securing the presence of at least one calm observer . |
| És az ott levő anyagból még a legügyesebb építész se tudta volna Éva lugasának az eszkimók hókunyhóinál különb mását felépíteni . | Nor , with such materials as were at hand , could the most skilful architect have constructed any better imitation of Eve 's bower than might be seen in the snow hut of an Esquimaux . |
| Az előkelő boltos igen nagyra értékelte a fürge és ügyes | All such proofs of a ready mind and skilful handiwork were highly |