META-NET: Language Technology for Europe

Georg Rehm

Network Manager META-NET
DFKI, Germany
georg.rehm@dfki.de

META-NET Roadshow Event – Bratislava, Slovakia
June 07, 2012

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

- Introduction
- META-SHARE
- Language White Paper Series
- Strategic Research Agenda
Challenge: Providing each language community with the most advanced technologies for communication and information so that maintaining their mother tongue does not turn into a disadvantage.

While research has made considerable progress in recent years, the pace of progress is not fast enough to meet the challenge within the next 10-20 years.

All stakeholders – researchers, LT user and provider industries, language communities, funding programmes, policy makers – should team up for a major dedicated push.
META-NET is a network of excellence dedicated to fostering the technological foundations of the European multilingual information society.

**META-VISION**: Building a community with a shared vision and strategic research agenda

**META-SHARE**: Building an open resource exchange infrastructure

**META-RESEARCH**: Building bridges to neighbouring technology fields
Four Funded Projects

- Initial project: T4ME (FP7; 13 partners, 10 countries)
- Three new consortia (ICT-PSP) since Feb. 2011: CESAR, METANET4U, META-NORD
- All EU member states and several non-member states covered.
- META-NET in June 2012: 57 members in 33 countries.

http://www.meta-net.eu
META-SHARE

META-SHARE
META-SHARE at a Glance

- Open exchange infrastructure for language resources and tools.
- Language resources and tools are documented, uploaded, stored in repositories, catalogued, can be downloaded, shared, discussed.
- Improve their visibility, documentation, identification, availability, preservation, interoperability.
- Long-term goal: boost research, technology and innovation through wide availability, pooling, openness and sharing of resources.
- Repositories store and maintain inventories of resources and tools.
- Metadata inventories are exported and harvested in the network.
- Currently 13 repositories up and running; ~1.300 LRs available.
META-SHARE

1,248 language resources at your disposal...

Type in your keywords, please...

Search  Browse  Statistics

What is it? - About the project

META-NET aims at creating 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...
JTok Tokenizer

 JTok Tokenizer
 JTok provides a tokenizer in Java that identifies paragraphs, sentences and tokens of an input text. Non-word tokens are further classified into abbreviations, numbers, punctuation and clitics.

 JTok

 http://heartofgold.opendfi.de/browser/trunk/jtok

 NOT_DEFINED_FOR_V2

 distributionInfo

 available-unrestrictedUse

 licenceInfo

 LGPL

 DistributionAccessMedium

 downloadable

 DownloadLocation

 http://heartofgold.opendfi.de/browser/trunk/jtok

 contactPerson

 Steffen

 Jörg

 male
Current Contents

Resource types

Distribu­tion per Language

- English: 503
- Spanish: 391
- French: 293
- German: 204
- Italian: 202
- Chinese: 199
- Dutch: 129
- Portuguese: 104
- Greek: 99
- Swedish: 76
- Catalan: 41
- Arabic: 41
- Japanese: 38
- Romanian: 36
- Danish: 30
- Polish: 28
- Other: 27

Language Description

Corpus
Lexical-Conceptual Resource
Tool Service

http://www.meta-net.eu
Recent Developments

- Several organizations outside META-NET want to set up their own META-SHARE repositories (e.g., JRC Optima Activity).
- Incoming resources from CESAR, METANET4U and META-NORD and also from additional initiatives (collaboration agreements with >40 EU-funded projects).
- ELRA will adopt META-SHARE as the central model for LR sharing and will provide services in its operation.
- META-SHARE Version 3.0 to be released in September 2012.

http://www.meta-net.eu
Key communication instruments to address decision makers and journalists.

Cover all EU languages (30 volumes).

Report on the state of a language (general, social, strategic and technological aspects) and the level of support through language technology.

Inform target group about societal and technological problems and challenges as well as economic opportunities.

Printed documents published by Springer; will be distributed by META-NET.

PDF versions available for free.
30 Languages Covered

- Basque
- Bulgarian*
- Catalan
- Czech*
- Danish*
- Dutch*
- English*
- Estonian*
- Finnish*
- French*
- Galician
- German*
- Greek*
- Hungarian*
- Icelandic
- Irish*
- Italian*
- Latvian*
- Lithuanian*
- Maltese*
- Norwegian
- Polish*
- Portuguese*
- Romanian*
- Serbian
- Slovak*
- Slovene*
- Spanish*
- Swedish*
- Croatian

* = Official EU language

http://www.meta-net.eu
THE SLOVAK LANGUAGE IN THE DIGITAL AGE

SLOVENSKÝ JAZYK V DIGITÁLNOM VÉKU

Mária Šimková
Radovan Garabik
Katarína Gajdošová
Michal Lacávnik
Slavomír Ondrejovič
Jozef Juhár
Ján Genči
Karol Furdič
Helena Tvariková
Jozef Ivanecký

Springer
1 Zhrnutie 1
2 Ohrozene našich jazykov: Výzva pre jazykové technológie 3
   2.1 Jazykové hranice spomáňajú európsku informačnú spoločnosť 4
   2.2 Naše jazyky v ohrození 4
   2.3 Jazykové technológie sú klúčovými technológiami 5
   2.4 Príležitosti pre jazykové technológie 5
   2.5 Výzvy pre jazykové technológie 6
   2.6 Osvojovanie s jazykami 6
3 Slovenčina v európskej informačnej spoločnosti 8
   3.1 Všeobecné lásky 8
   3.2 Špecifika slovenčiny 11
   3.3 Slovenčina na internete 12
   3.4 Slovenčina ako cudzí jazyk 13
   3.5 Slovenský národný korpus 15
4 Jazykové technológie na podporu slovenčiny 17
   4.1 Architektúra aplikácií 17
   4.2 Základné aplikáčné oblasti 19
   4.3 Ďalšie aplikáčné oblasti 27
   4.4 Jazykové technológie vo vzdelení 29
   4.5 Státné programy a iniciatívy 29
   4.6 Dostupnosť nástrojov a zdrojov 30
   4.7 Porovnanie jazykov 32
   4.8 Závery 33
5 O META-NETe 37
A few Numbers ...

- >160 national experts contributed as authors or co-authors (ca. 5 per language on average)
- >50 additional experts have contributed data and information
- >8,000 copies will be printed and distributed by META-NET

http://www.meta-net.eu
Cross-Lingual Ranking

- Journalists and politicians need simple and clear messages.
- In four application areas, each language is assigned to one of five clusters, ranging from excellent LT support to weak/no support:
  1. Machine Translation
  2. Speech Processing
  3. Text Analysis
  4. Resources
- Results finalised at a meeting in Berlin with representatives of all 30 languages (October 21/22, 2011).

http://www.meta-net.eu
<table>
<thead>
<tr>
<th>excellent</th>
<th>good</th>
<th>moderate</th>
<th>fragmentary</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>English</td>
<td>French, Spanish</td>
<td>Catalan, Dutch, German, Hungarian, Italian, Polish, Romanian</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Basque, Bulgarian, Croatian, Czech, Danish, Estonian, Finnish, Galician, Greek, Icelandic, Irish, Latvian, Lithuanian, Maltese, Norwegian, Portuguese, Serbian, Slovak, Slovene, Swedish</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>Czech, Dutch, Finnish, French, German, Italian, Portuguese, Spanish</td>
<td>Basque, Bulgarian, Catalan, Danish, Estonian, Galician, Greek, Hungarian, Irish, Norwegian, Polish, Serbian, Slovak, Slovene, Swedish</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Croatian, Icelandic, Latvian, Lithuanian, Maltese, Romanian</td>
</tr>
<tr>
<td>excellent</td>
<td>good</td>
<td>moderate</td>
<td>fragmentary</td>
</tr>
<tr>
<td>-----------</td>
<td>------</td>
<td>----------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>English</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Dutch, French, German, Italian, Spanish</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Basque, Bulgarian, Catalan, Czech, Danish, Finnish, Galician, Greek, Hungarian, Norwegian, Polish, Portuguese, Romanian, Slovak, Slovene, Swedish</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Czech, Dutch, French, German, Hungarian, Italian, Polish, Spanish, Swedish</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Basque, Bulgarian, Catalan, Croatian, Danish, Estonian, Finnish, Galician, Greek, Norwegian, Portuguese, Romanian, Serbian, Slovak, Slovene</td>
</tr>
</tbody>
</table>
Europe’s Languages and LT

- English
- Dutch
- French
- German
- Italian
- Spanish
- Catalan
- Basque
- Bulgarian
- Czech
- Danish
- Finnish
- Glovesian
- Hungarian
- Greek
- Hungarian
- Polish
- Portuguese
- Romanian
- Swedish
- Slovak
- Slovene
- Croatian
- Estonian
- Bulgarian
- Croatian
- Greek
- Icelandic
- Irish
- Latvian
- Lithuanian
- Maltese
- Romanian
- Serbian
- Slovene
- Swedish

**good support through Language Technology**

**weak or no support**

http://www.meta-net.eu
Not enough R&I on European languages, except for English, is too weak and too slow. Many languages are badly covered.

Languages treated in the 2010 editions of Journal of Computational Linguistics and Conferences of ACL, EMNLP and COLING. Many European languages without any reference: Slovak, Maltese, Lithuanian, Irish, Albanian, Croatian, Galician etc.
Observations

- When it comes to Language Technology support, there are massive differences between Europe’s languages and technology areas.
- LT support for English is ahead of any other language – but far from being perfect.
- For 16 of 30 languages, LT support is only fragmentary or very weak!

META-VISION

Strategic Research Agenda

http://www.meta-net.eu
Mobilize researchers, users and providers of LT for cooperation and collaboration to initiate and support a large-scale joint action to

- Building a **community** around Language Technology in Europe (META),
- Creating a **shared vision**, and
- Preparing a **Strategic Research Agenda** for Multilingual Europe 2020, to be presented to national/international politicians, funding agencies, decision makers.
From Visions to the SRA

- Three **Vision Groups** brought together researchers, developers, integrators and (corporate or professional) users of LT-based products, services and applications (ca. 25 members each).

- Collected domain-specific visions and prepared individual reports.

- Vision Group **Translation and Localisation**
  - July 23, 2010, Berlin, Germany
  - September 28, 2010, Brussels, Belgium
  - April 7/8, 2011, Prague, Czech Republic

- Vision Group **Media and Information Services**
  - September 10, 2010, Paris, France
  - October 15, 2010, Barcelona, Spain
  - April 1, 2011, Vienna, Austria

- Vision Group **Interactive Systems**
  - September 10, 2010, Paris, France
  - October 5, 2010, Prague, Czech Republic
  - March 28, 2011, Utrecht, The Netherlands
Steps taken towards the SRA

- Many suggestions by: Vision Group members; META-NET; META.
- Additional input in meetings, workshops, discussions etc.
- Language White Papers: research gaps; funding situation etc.
- We screened the Strategic Research Agendas of other initiatives.
- We discussed procedures, input and structure of the SRA in three meetings of the META Technology Council.
  - Brussels, Belgium, November 16, 2010
  - Venice, Italy, May 25, 2011
  - Berlin, Germany, September 30, 2011
  - Upcoming: Brussels, Belgium, June 19, 2012
From Visions to the SRA
Priority Research Themes

Translation Cloud
Translation Technologies
Core Technologies for Language Analysis & Production
Technologies for Knowledge Discovery and Analysis
Social Intelligence
Technologies for Speech Recognition and Analysis
Socially Aware Interactive Assistant

LT 2020
Vision and Priority Themes for Language Technology Research in Europe until the Year 2020
Towards the META-NET Strategic Research Agenda

The development of this paper has been funded by the Seventh Framework Programme and the ICT Policy Support Programme of the European Commission under contracts T4ME (Grant Agreement 249119), CESAR (Grant Agreement 271022), METANET4U (Grant Agreement 270893) and META-NORD (Grant Agreement 270899).

Do you have comments, ideas or suggestions with regard to the content of this document? Please send them to office@meta-net.eu or discuss them online: http://www.meta-net.eu/sra.
META-NEG is a network of excellence.

META is an open and growing strategic technology alliance: Multilingual Europe Technology Alliance.

- 600+ members, including W3C, Google, Microsoft, GALA, research centres, LT companies etc.
- Main goal: to support our Strategic Research Agenda.
- Join us!  http://www.meta-net.eu/join
language communities

policy makers and funding bodies

user industries

provider industries

language technology community

semantic technologies community

language technology community

machine learning community

multimedia content technologies

cognitive systems community

t4me METANET4U

META COUNCIL

META NET

META NORD

META

ISO

W3C

EUROPEAN PARLIAMENT

GALA

Globalization & Localization Association

FNIL

LT-INNOVATE.EU
Next Steps

- Disseminate the Language White Papers widely.
- Meet with national research planners, funders, policy makers and inform them about the Strategic Research Agenda.
- Mobilize user industries and administrations.
- Upcoming opportunities can provide sufficient resources to make our visions for Europe’s citizens and economy, as described in the SRA, a reality (Horizon 2020 and Connecting Europe Facility, CEF).

http://www.meta-net.eu
Thank you very much!

See you in Brussels on June 20/21 for META FORUM 2012
http://www.meta-forum.eu

office@meta-net.eu
http://www.meta-net.eu
http://www.facebook.com/META.Alliance