The META-NET White Paper Series: Languages in the European Information Society

Georg Rehm

DFKI, Germany
georg.rehm@dfki.de

EFNIL 9th Annual Conference – London, UK
October 26, 2011

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-NET and META
- The META-NET Language White Paper Series
- Towards a Strategic Research Agenda for Multilingual Europe

Multilingual Europe

META-NET

- **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**.

Objectives

META-NET is a network of excellence dedicated to fostering the technological foundations of the European multilingual information society.

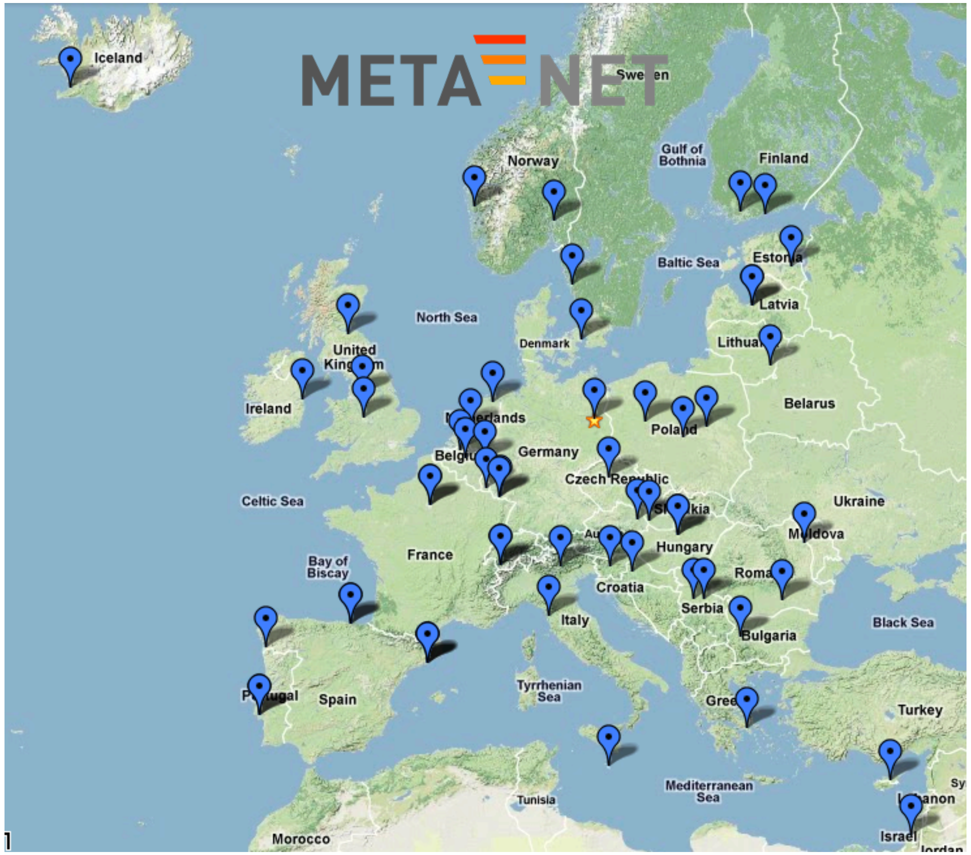
META-VISION: Building a community with a shared vision and strategic research agenda

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 support consortia (ICT-PSP) since Feb. 2011.
- All EU member states and several non-member states covered.
- META-NET in October 2011: **54** members from **33** countries.



http://www.meta-net.eu/members

META-NET (October 21, 2011)

META-NET Network Meeting and General Assembly, October 21/22, 2011, Berlin, Germany

META

- **META-NET** is a network of excellence.
- **META** is an open and growing strategic technology alliance: **Multilingual Europe Technology Alliance**.
  - Almost 300 members, including W3C, Google, Microsoft, GALA, research centres, LT companies etc.
  - **META** includes multiple stakeholders to prepare the ground for a large-scale concerted effort.
  - Main goal: to support our Strategic Research Agenda.
  - Join us! **http://www.meta-net.eu/join**

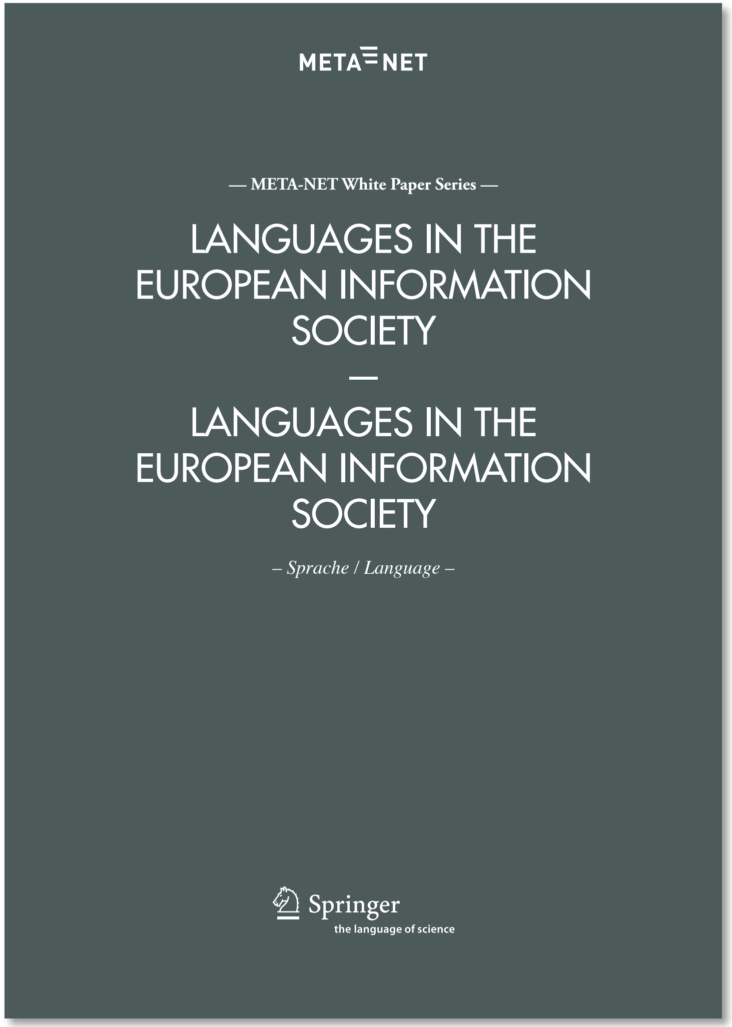
META-VISION

The META-NET Language White Paper Series

The Language White Papers

META-NET

- LT support varies massively from language to language.
- Raise awareness for the topic; inform about the current status.
- Survey of the state of 30 languages in the digital society.
- Target audience: national and international politicians, journalists, decision makers, the public at large.
- Key messages: societal and technological problems, challenges, economic opportunities.



META-NET

— META-NET White Paper Series —

LANGUAGES IN THE EUROPEAN INFORMATION SOCIETY

—

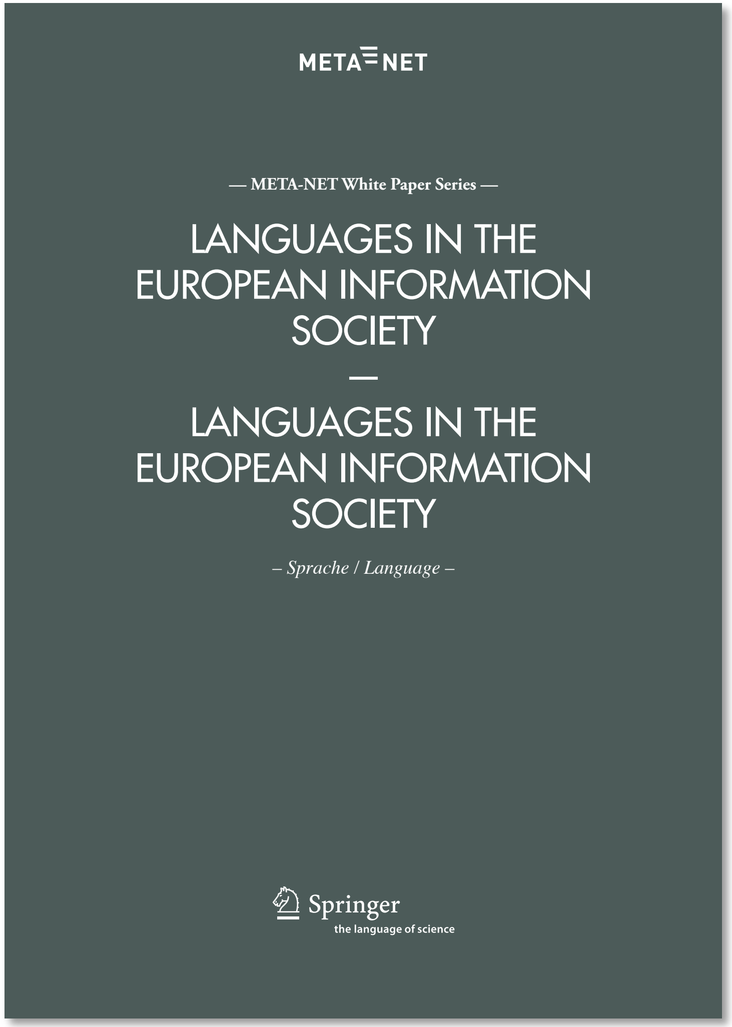
LANGUAGES IN THE EUROPEAN INFORMATION SOCIETY

– Sprache / Language –

Springer
the language of science

Structure

- Part 1: Introduction – A Risk for Our Languages and a Challenge for LT
- Part 2: *Language* in the European Information Society
- Part 3: LT Support for *Language*
- Part 4: Conclusions and Cross-Language Comparison



META-NET

— META-NET White Paper Series —

LANGUAGES IN THE EUROPEAN INFORMATION SOCIETY

—

LANGUAGES IN THE EUROPEAN INFORMATION SOCIETY

– *Sprache / Language* –

Springer
the language of science

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

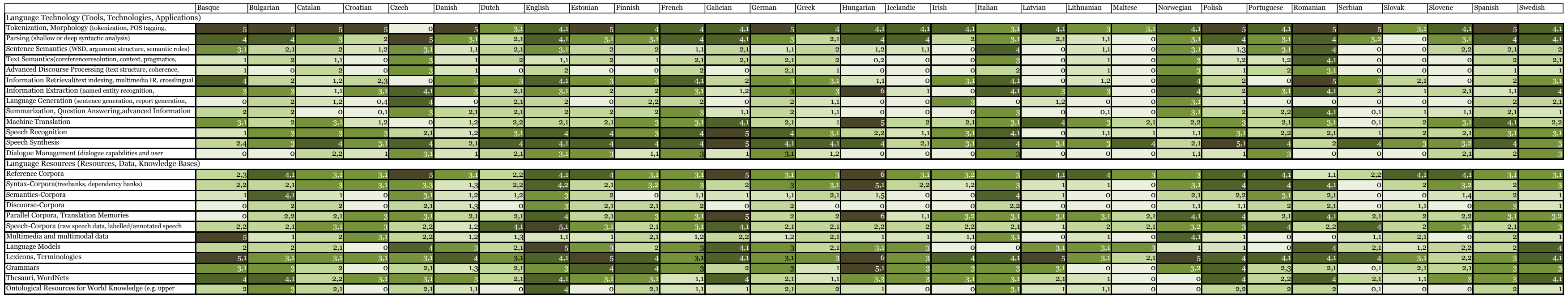
* = Official EU language

How to Assess LT Support?

- How to assess LT support for a certain language? How to arrive at results that can be communicated to journalists?
  - Count tools and resources? *Message wouldn't really be meaningful.*
  - Define quality criteria and perform a comparative evaluation? *Complicated, complex, time-consuming process – would take years.*
- Solution: experts provided estimations condensed in a table assessing core areas such as MT, IR, ASR, parsing, corpora etc.
- >150 national experts contributed (ca. 5 per language on avg.).
- Seven assessment criteria: *availability*, *quality*, *quantity*, *coverage*, *maturity*, *sustainability* and *adaptability*.

Intermediate Steps

- 30 tables provide data for all languages (tools, resources, gaps etc.).
- Reduce numbers to one final score per language and area.
- Calibration of tables across languages in smaller groups.
- Final scores for each area and language were derived from the two central features (quality, coverage), resulting in *one big* table:



	Basque	Bulgarian	Catalan	Croatian	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
Language Technology (Tools, Technologies, Applications)																														
Tokenization, Morphology (tokenization, POS tagging,	5	5	5	5	0	5	3,1	4,1	5	4	4	4,1	5	4	4,1	4,1	4,1	3,1	4,1	3	3,1	4,1	5	4,1	5	5	3,1	4,1	5	4,1
Parsing (shallow or deep syntactic analysis)	4	4	3	2	5	3,1	2,1	4,1	3,1	3,1	4	4,1	3	2,1	4	4	2	3,1	2,1	1,1	0	3,1	4	3,1	4	3,2	0	3,1	4	4,1
Sentence Semantics (WSD, argument structure, semantic roles)	3,1	2,1	2	1,2	3,1	1,1	2,1	3,1	2	2	1,1	2,1	1,1	2	1,2	1,1	0	4	0	1,1	0	3,1	1,3	3,1	4	0	0	2,2	2,1	2
Text Semantics(coreferenceresolution, context, pragmatics,	1	2	1,1	0	3	1	2	1,1	2	1	2,1	2,1	2,1	2	0,2	0	0	3	0	1	0	3	1,2	1,2	4,1	0	0	0	2	2,1
Advanced Discourse Processing (text structure, coherence,	1	0	2	0	3	1	0	2	0	0	2	0	2,1	1	0	0	0	2	0	1	0	3	1	2	3,1	0	0	0	1	1
Information Retrieval(text indexing, multimedia IR, crosslingual	4	2	1,2	2,3	0	3	3	4,1	3	3	4,1	2	3	3,1	1,1	0	3,1	4,1	0	1,2	0	4	2	0	5	3	2,1	0	2	3,1
Information Extraction (named entity recognition,	3	3	1,1	3,1	4,1	3	2,1	3,1	2	2	3,1	1,2	3	3	6	1	0	4,1	3	3	0	4	2	3,1	4,1	2	1	2,1	1,1	4
Language Generation (sentence generation, report generation,	0	2	1,2	0,4	4	0	2,1	2	0	2,2	2	0	2	1,1	0	0	3	0	1,2	0	0	3,1	1	0	0	0	0	0	2	2,1
Summarization, Question Answering,advanced Information	2	2	0	0,1	3	2,1	2,1	2	2	2	3	1,1	2	1,1	0	0	0	3	0	0,1	0	3,1	2	2,2	4,1	0,1	1	1,1	2,1	1
Machine Translation	3,1	2	3,1	1,2	0	1,2	2,2	2,1	2,1	3	3,1	4,1	2,1	1	5	2	2,1	3,1	4	3	2,1	2,2	3	2,1	3,1	0,1	2	3,1	4,1	2,2
Speech Recognition	1	3	3	3	2,1	1,2	3,1	4	4	3	4	5	4	3,1	2,2	1,1	3,1	4,1	0	1,1	1	1,1	3,1	2,2	2,1	1	2	2,1	3,1	3,1
Speech Synthesis	2,4	3	4	3,1	4	2,1	4	4,1	4	4	4	5	4,1	4,1	4	2,1	3,1	4	3,1	3	4	2,1	5,1	4	2	4	3	3,2	4	3
Dialogue Management (dialogue capabilities and user	0	0	2,2	1	3,1	1	2,1	3,1	3	1,1	3	1	3,1	1,2	0	0	0	3	0	0	0	1,1	1	3	0	0	0	2,1	2	3
Language Resources (Resources, Data, Knowledge Bases)																														
Reference Corpora	2,3	4,1	3,1	3,1	5	3,1	2,2	4,1	4	3,1	3,1	5	3,1	3	6	3,1	3,2	3	4,1	4	3	3	4	4,1	1,1	2,2	4,1	4,1	3,1	3,1
Syntax-Corpora(treebanks, dependency banks)	2,2	2,1	3	3,1	3,3	1,3	2,2	4,2	2,1	3,2	3	2	3	3,1	5,1	2,2	1,2	3	1	1	0	3,1	4	4	4,1	0	2	3,2	2	3
Semantics-Corpora	1	4,1	1	0	3,1	1,2	1,2	3	2	0	1,1	1	1,1	2,1	1,5	0	0	4	1	0	0	2,1	2,2	3,1	2,1	0	0	1,4	2	1
Discourse-Corpora	0	2	2	0	2,1	1,3	0	3	2,1	2,1	2	0	2	0	0	0	0	2,2	0	0	0	1,1	1,1	2	2,1	0	1,1	0	3	1
Parallel Corpora, Translation Memories	0	2,2	2,1	3	3,1	2,1	2,1	4	2,1	3	3,1	5	2	2	6	1,1	3,2	3,1	3,1	3,1	2,1	4,1	4	2,1	4,1	2,1	2	2,2	3,1	3,2
Speech-Corpora (raw speech data, labelled/annotated speech	2,2	2,1	3,1	3	2,2	1,2	4,1	5,1	3,1	2,1	3,1	4,1	2,1	2,1	2,2	2	2,2	2,1	1	2	2,1	3,2	3	4	2,2	4	2	3,1	2,1	3
Multimedia and multimodal data	5	1	2	3,1	2,2	1,2	1,3	1,1	1	2,1	1,2	2,2	1,2	2,1	1	1	1,1	3,1	0	1	0	4,1	1	0	0	1,1	2,1	0	2	1
Language Models	2	2	2,1	0	4	3	2,1	5	3	2	3	4,1	3	2,1	3,1	3	0	0	3,1	3,1	3	1	1	0	4	2,1	1,2	2,2	2	4
Lexicons, Terminologies	5,1	3,1	3,1	3,1	3,1	4	3,1	4,1	5	4	3,1	4,1	3,1	3	6	3	4	4,1	5	3,1	2,1	5	4	4,1	4,1	4	3,1	2,2	3	4,1
Grammars	3,1	3	2	0	2,1	1,3	2,1	3	4	4	3	2	3	1	5,1	3	3	3	3,1	0	0	3,2	4	2,3	2,1	0,1	2,1	2,1	3	3
Thesauri, WordNets	4	4,1	2,2	3,1	3,1	3	2,1	4,1	3,1	3,1	1,1	4	2,1	1,1	3,3	3	3,1	3,1	2,1	1	0	0	4	2,2	4	2,1	1,1	3	3	4,1
Ontological Resources for World Knowledge (e.g. upper	2	3	2,1	0	2,1	1,1	0	4	0	2,1	1,1	1	2,1	2	1	0	0	3,1	1	1,1	0	0	2,2	2	2	0,1	0	0	2	1

Cluster-Based Presentation

- For journalists and politicians the big table is useless.
- Therefore: Cluster-based cross-language comparison
- Each language is assigned to one of five clusters, ranging from *excellent LT support* to *weak/no support*.
- Presentation of key results with regard to four areas:
  - Area 1: Machine Translation
  - Area 2: Speech Processing
  - Area 3: Text Analysis
  - Area 4: Resources
- Clusters discussed and finalised at a recent meeting in Berlin with representatives of all 30 languages.

MT (top) & Speech Processing (bottom)

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
	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

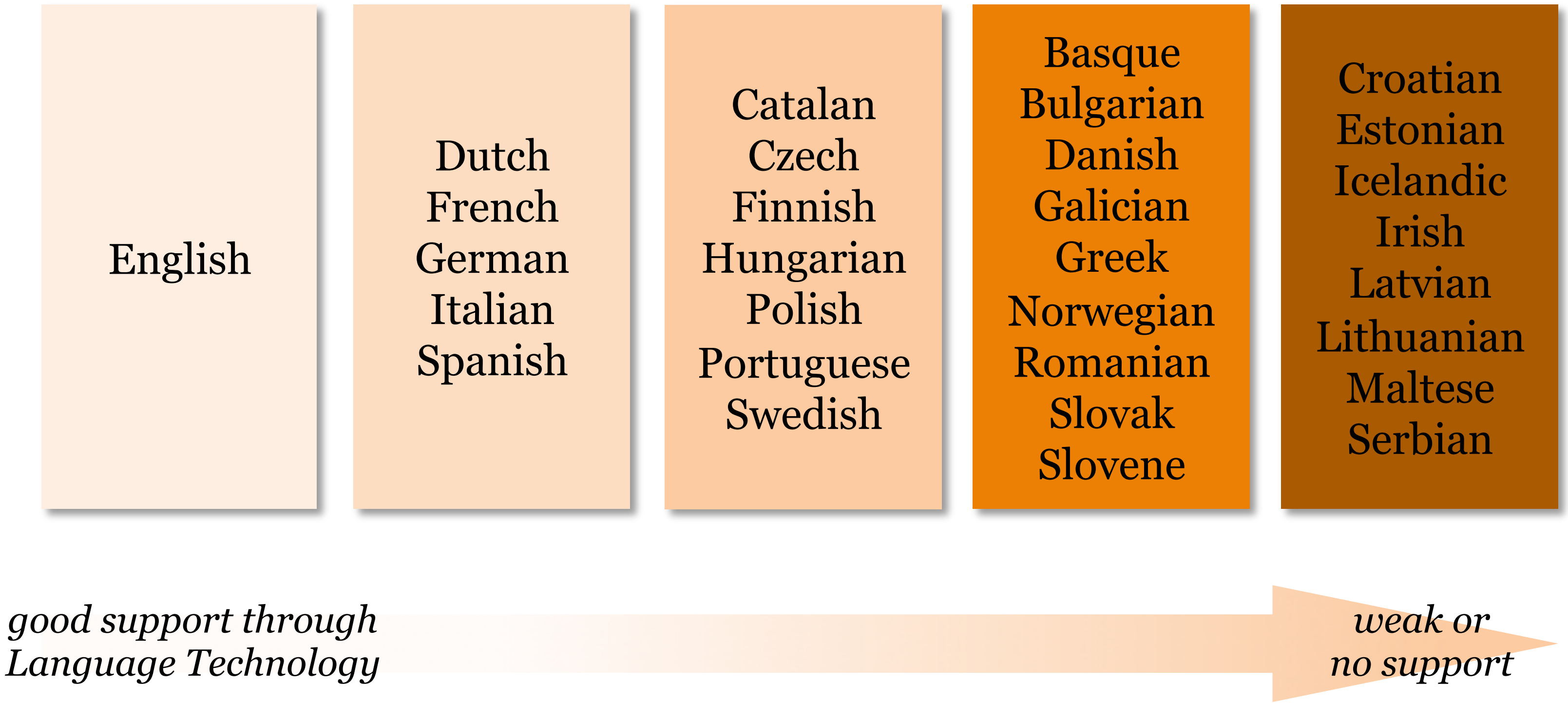
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
	English	Czech, Dutch, Finnish, French, German, Italian, Portuguese, Spanish	Basque, Bulgarian, Catalan, Danish, Estonian, Galician, Greek, Hungarian, Irish, Norwegian, Polish, Serbian, Slovak, Slovene, Swedish	Croatian, Icelandic, Latvian, Lithuanian, Maltese, Romanian

Text Analysis (top) & Resources (bottom)

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
	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

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
	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

Europe’s Languages and LT



English

Dutch
French
German
Italian
Spanish

Catalan
Czech
Finnish
Hungarian
Polish
Portuguese
Swedish

Basque
Bulgarian
Danish
Galician
Greek
Norwegian
Romanian
Slovak
Slovene

Croatian
Estonian
Icelandic
Irish
Latvian
Lithuanian
Maltese
Serbian

good support through
Language Technology

weak or
no support

Observations

META NET

- When it comes to Language Technology support, there are massive differences between languages and technology areas.
- For all LT areas, English is ahead of any other language.
- Even support for English is far from being perfect.
- For 16 languages, LT support is only fragmentary or very weak.
- Most (very) large companies have stopped working in LT, leaving the field to SMEs, which can hardly attack an international market.

- Draft versions available at http://www.meta-net.eu/whitepapers
- Final printed white papers to be available in early 2012.

META-NET

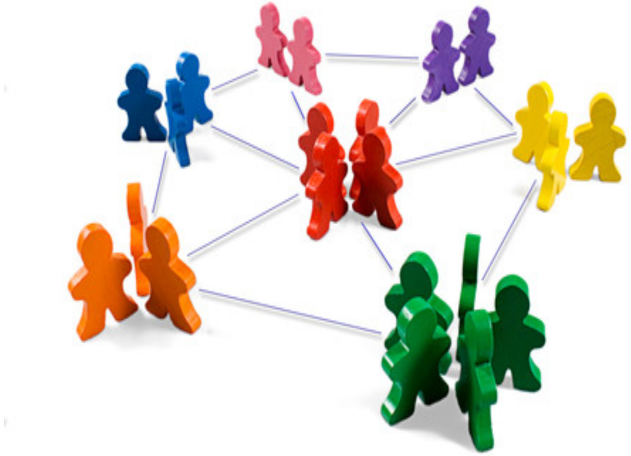
META-VISION

Towards a Strategic Research Agenda for Multilingual Europe

Shared Vision and SRA

META NET

- Mobilize researchers, decision makers, users and providers of LT, R&D programmes for cooperation and collaboration in META.
- Large-scale joint action to
  - Building a community around Language Technology in Europe (META)
  - **Creating a shared vision**
  - **Preparing a Strategic Research Agenda for Multilingual Europe**



From Visions to the SRA

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

- They collect domain-specific visions and prepare 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

From Visions to the SRA

- **META Technology Council** prepares two documents:
  - **The Future European Multilingual Information Society. Vision Paper for a Strategic Research Agenda**
    http://www.meta-net.eu/vision/reports/meta-net-vision-paper.pdf
  - **Strategic Research Agenda for Multilingual Europe.**
    - To be presented to national and international politicians, administrators and funding agencies.
    - Will cover a timeframe from now to ca. 2025
    - *Work in progress.*

www.meta-net.eu
office@meta-net.eu
T: +49 30 23895 1833

The Future European Multilingual Information Society

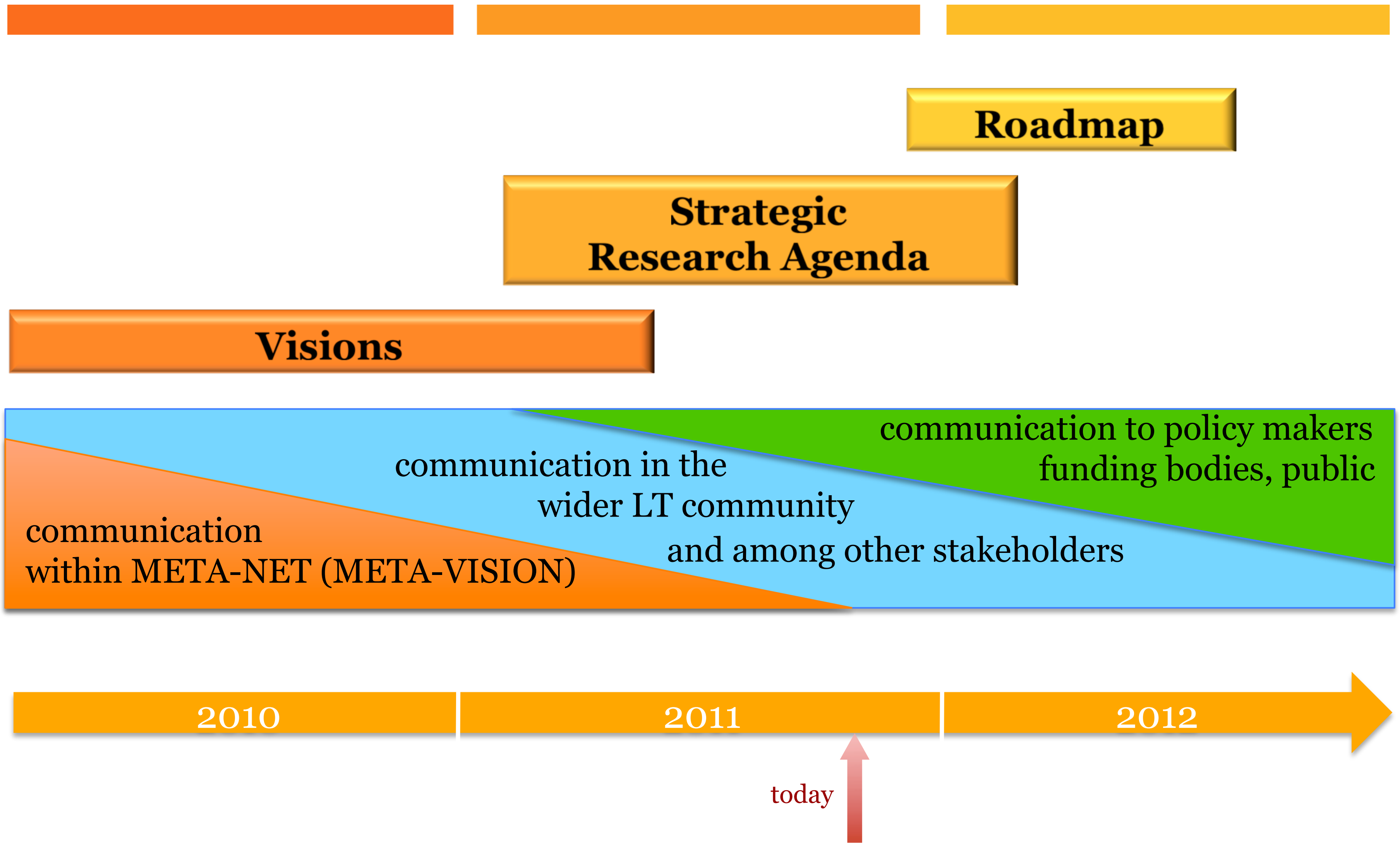
Vision Paper for a Strategic Research Agenda

"People can't share knowledge
if they don't speak a common language."

Davenport, Thomas H, and Laurence Prusak, *Working Knowledge: How Organizations Manage What They Know*, Harvard Business School, Boston, 1997, p. 98.

Join the discussion at
www.meta-et.eu/forum

The Planning Process



Roadmap

Strategic Research Agenda

Visions

communication to policy makers funding bodies, public

communication in the wider LT community and among other stakeholders

communication within META-NET (META-VISION)

2010

2011

2012

today

Towards the SRA

- Many suggestions by the Vision Group members.
- Additional input in meetings, workshops, discussions 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

SRA: Structure

Letter from the META-NET Partners

META: Multilingual Europe Technology Alliance

1. Executive Summary
2. Multilingual Europe: Facts, Challenges, Opportunities
3. ICT: Current State, Major Trends and Predictions
4. Language Technology: State, Limitations, Potential
5. Language Technology for Multilingual Europe: The Grand Vision
6. Language Technology for Multilingual Europe: Priorities, Plans, Roadmap

References

List of Contributors

Lead Vision Candidates

- Translation Cloud – Translation Everywhere, Everytime
- Talking Things – Talking Everything – Wised-Up World
- Multilingual e-Participation – Translingual e-Democracy
- Crosslingual e-Learning
- Virtual Avatar – Second me – Talking Assistant
- Talking Robots
- Intelligent Information Organiser

Get Involved!

- The **Multilingual Europe Technology Alliance** needs as many members as possible.
- A few EFNIL members are already members of META (such as, for example, Dansk Sprognævn).
- If you're not a member, please join us!

http://www.meta-net.eu/join

– no financial obligations whatsoever –

- Please also talk to your industry contacts, approach your friends and colleagues at universities, research centres, institutes, companies, startups! Ask them to join **META**!
- Get involved and have your say!

Q/A

Thank you very much!

Joint work with Aljoscha Burchardt, Kathrin Eichler, Felix Sasaki, Hans Uszkoreit (all from DFKI), the ca. 70 members of the Vision Groups, the 30 members of the META Technology Council and the ca. 150 authors of and contributors to the META-NET Language White Papers.

office@meta-net.eu

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