Human Language Technology Days 2012

META-NET: building the technological foundations of the multilingual European information society

Hans Uszkoreit, DFKI

Coordinator META-NET

Multilingual Europe

- Egalitarian multilingual society of the European Union is an ambitious endeavour and an unprecedented socioeconomic experiment.
- Two dozen national and many regional languages (total > 40)
- One core component is a common market with a single information space.

Our last borders…

… are language borders

→ After removing barriers for people, goods and capital, barriers still exist for the free fow of thought, knowledge, creative content, and other information.

→ After the Fukushima accident, nuclear energy was discussed in social fora throughout Europe – but never across language borders

Our last borders…

… are language borders

→ After removing barriers for people, goods and capital, barriers still exist for the free fow of thought, knowledge, creative content, and other information.

→ After the Fukushima accident, nuclear energy was discussed in social fora throughout Europe – but never across language borders

→ granice językowe

The Role of Technology

META Multilingual Europe Technology Alliance

language days

- IT (especially Internet and language technology) is part of the problem…
- …but it is also the source of the solution

Major Challenges

- Preserving the European cultural and linguistic diversity in the united information and knowledge society
- Securing at affordable costs the free flow of information and thought across language boundaries in the resulting single information space
- Providing each language community with the most advanced technologies for communication, information and knowledge management so that maintaining their mother tongue does not turn into a disadvantage

Today LT is already surrounding us

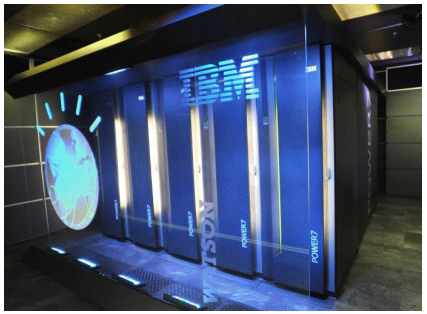
META Multilingual Europe Technology Alliance

language days

- spell/grammar checker in MS Word
- voice dialing on the cellphone
- web search in Google
- speech generation in game software
- computer-assisted language learning
- optical character recognition
- semantic text classification in Autonomy
- speech control in cars
- voice dialogues in call centers

We are witnessing the breakthrough of LT

- UK Text Analytics Company Autonomy bought for 8bEUR by HP
- IBM Watson wins Jeopardy
- Google renames its Division “Search” to “Knowledge”
- Siri improves the iPhone
- Google Translate covers 57 languages
- All large IT corporations, EC, EP and EPO deploy new generation translation technology



The Downside

But then we still cannot…

…translate the full meaning of any nontrivial paragraph into a MRL

or

…reliably translate a journal article

But this is only the beginning…

META Multilingual Europe Technology Alliance

language days

… since LT is a key enabling technology such as network technology, database technology or web technology

→ it is just much more complex because of the size of language (words, expressions, constructions, variants of language, and number of languages).

Why Key Enabling Technology?

META Multilingual Europe Technology Alliance

language days

- LT will overcome communication barriers
  - between people and technology,
  - between people speaking different languages

- and LT will unleash the full power of IT
  - for managing and better utilizing humankind's accumulated knowledge,
  - for producing, managing and accessing creative content,
  - for effectively mastering and exploiting the never-ending explosion of newly created information.

- The Internet is the medium that can overcome the language barriers and the support problem
- The Internet also offers business opportunities for numerous SME LSPs (language service providers)
- Translation, text analytics and speech recognition/production for all surviving languages will be offered as cloud-based services
- European Research is working with the W3C to make the next generation of the Web truly multilingual (Projects “Multilingual Web” and “LT-Web”)
- Long-Term Vision: The Cross-lingual (Translingual) Semantic Web

European Factors

- European institutions also have a growing demand for language technology
- EC DGT is using machine translation from our EU funded projects
- The European Parliament is building up similar solutions
- Many other European institutions start following
- European Patent Office turned to Google for faster help

Its current markets are big

META Multilingual Europe Technology Alliance

language days

- 24 BEuro worldwide speech-technology products and services,
- 20 BEuro worldwide translation products and services;
- 50% of the market in Europe;
- 500.000 translation/language professionals in Europe,
- annual growth 10-13%
  (much higher than general economy)
- Similar figures in markets for text analytics, language learning, language proofing, media subtitling/captioning, etc.

The future markets …

… are only limited by the number of people on earth, the number of their ICT devices, used services, the volume of written knowledge, written and spoken content, and all other information expressed in language.

The demand for LT is growing fast…

META Multilingual Europe Technology Alliance

language days

… because of several factors:

- → globalization (e-commerce and mobility by tourism and migration)
- → explosion of knowledge, creative content and other information
- → spread of advanced technology into all geographic regions and all parts of society (Internet, mobile communication, automobiles, consumer electronics, in the near future also ubiquitous services, smart homes and service robots).

Two important factors for us in Europe…

- is European integration with the legal and political obligations following from the egalitarian and inclusive approach to the languages and cultures of its member states.
- EU markets are multilingual … but so are our export markets.

European LT research is strong

META Multilingual Europe Technology Alliance

- EU research has achieved many important advances in MT and other areas.
- We are competing successfully with US and Asian research
- We have managed to get machine translation to the users

- **But considering the number and complexity of languages and applications, research is spotty and underfunded**

European language industry is big

META Multilingual Europe Technology Alliance

language days

- → Thousands of language service enterprises
  translation, interpretation, authoring, language teaching
- → Hundreds of IT companies with LT products

but it is fragmented

- → Almost excelusively SMEs
- → Suffering from lack of coordination, standards, interoperability.

Europe has greater demand

META Multilingual Europe Technology Alliance

language days

- LT is an area in which Europe has a greater demand than its main competitors, a greater potential but also much greater opportunities.
- In Europe LT addresses at the same time recognized societal needs (inclusion, single digital space, linguistic and cultural diversity) and opens an opportunity for business in a growth area in which we have a clear competitive advantage.
- After having missed the lead in several key enabling technologies Europe has the chance to come out ahead in this key enabling software technology.

Reality is different

- Unfortunately, today reality looks different. Europe is loosing talents to other parts of the world
- The main figures behind Google Translate, LocalizationWorld, Trados, etc. are mainly Europeans.
- Europe is also loosing intellectual outcome of successful research to commercialization in other parts of the world
  - by migration of talent,
  - uptake abroad
  - acquisition of start-ups that do not have the needed venture capital and other support for thriving.

Not enough R&I on European languages

META Multilingual Europe Technology Alliance

language days

- → LT research on European languages, except for English, is too weak and too slow.
- → Many languages are badly covered.

The Language White Papers

META Multilingual Europe Technology Alliance

language days

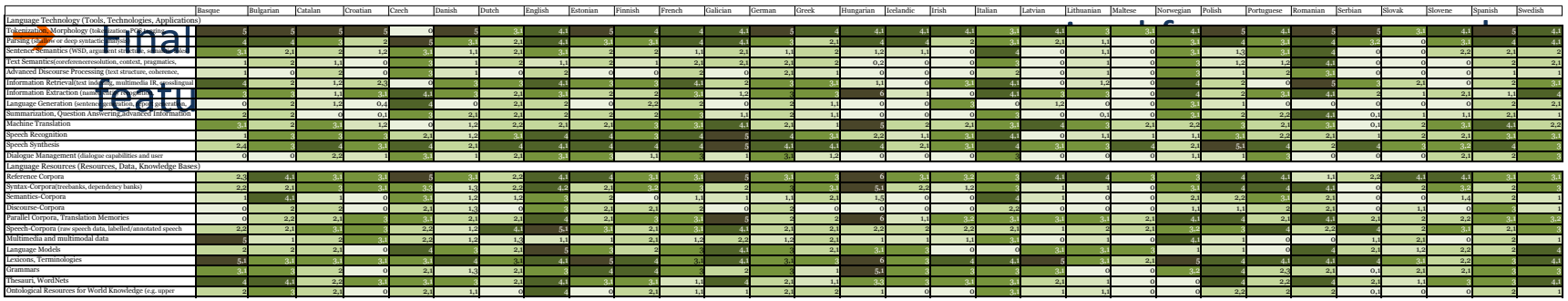
- >2 years in the making.
- >200 national experts as authors, co-authors or contributors (ca. 7 per language on average)
- >8.000 printed copies will be printed and distributed by META-NET to politicians and journalists.

Simplified Methodology

META Multilingual Europe Technology Alliance

language days

- Distributed data collection process in the respective countries.
- 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.



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

Cross-Lingual Ranking

- 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).

MT (top) & Speech Processing (bottom)

META Multilingual Europe Technology Alliance

language days

Machine Translation

excellent	good	moderate	fragmentary	weak or no support
	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

Speech Processing

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

Text Analysis (top) & Resources (bottom)

META Multilingual Europe Technology Alliance

language days

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

Language 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

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

- 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.
- Even support for English is *far* from being perfect.
- The gap between English and the other languages keeps widening!
- Several languages – such as Icelandic, Latvian, Lithuanian, Maltese – receive this weakest score in all four areas!
- *At least 21 European languages in danger of digital extinction* (languages in the "weak or no support" category for some area)!

Findings of 30 Language White Papers

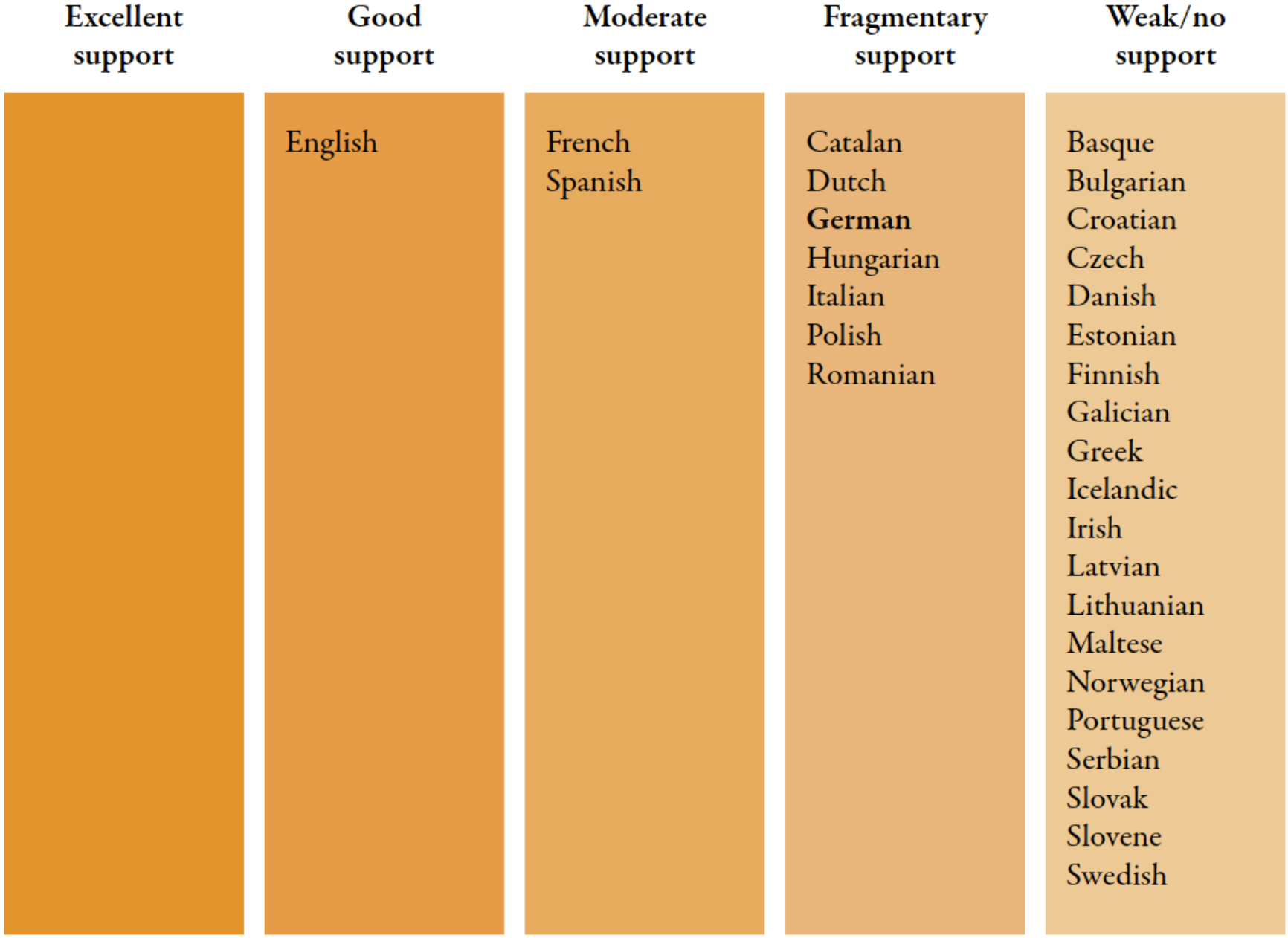
META Multilingual Europe Technology Alliance

language days

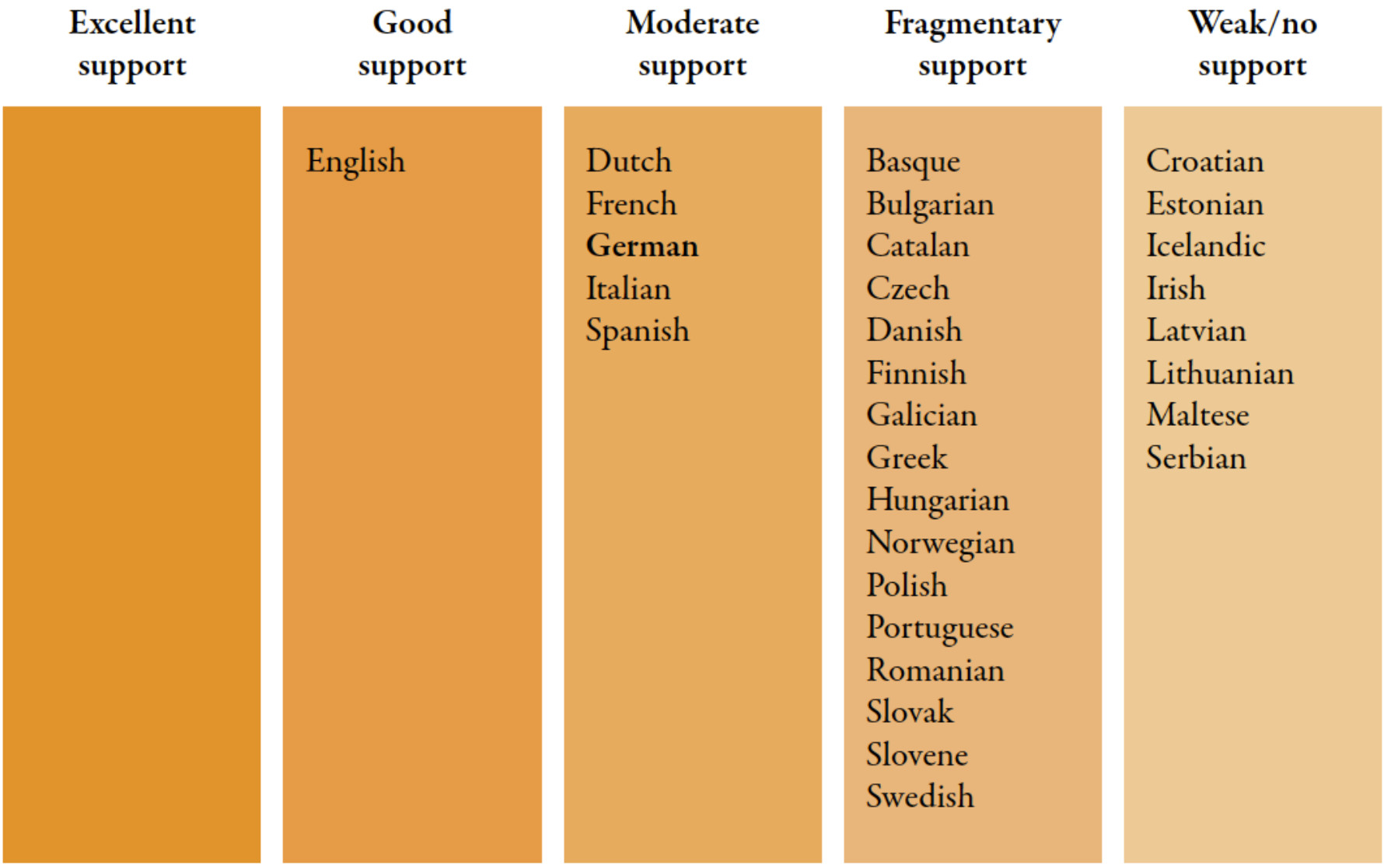
- In our 30 Language White Papers, we have surveyed the state of each language with respect to its status and technological supprt in the digital age.

- The observed differences are immense. Many European languages are severely under-supported.

- At the current level of research and technology development, the gap keeps widening year by year.



10: Machine translation: state of language technology support for 30 European languages



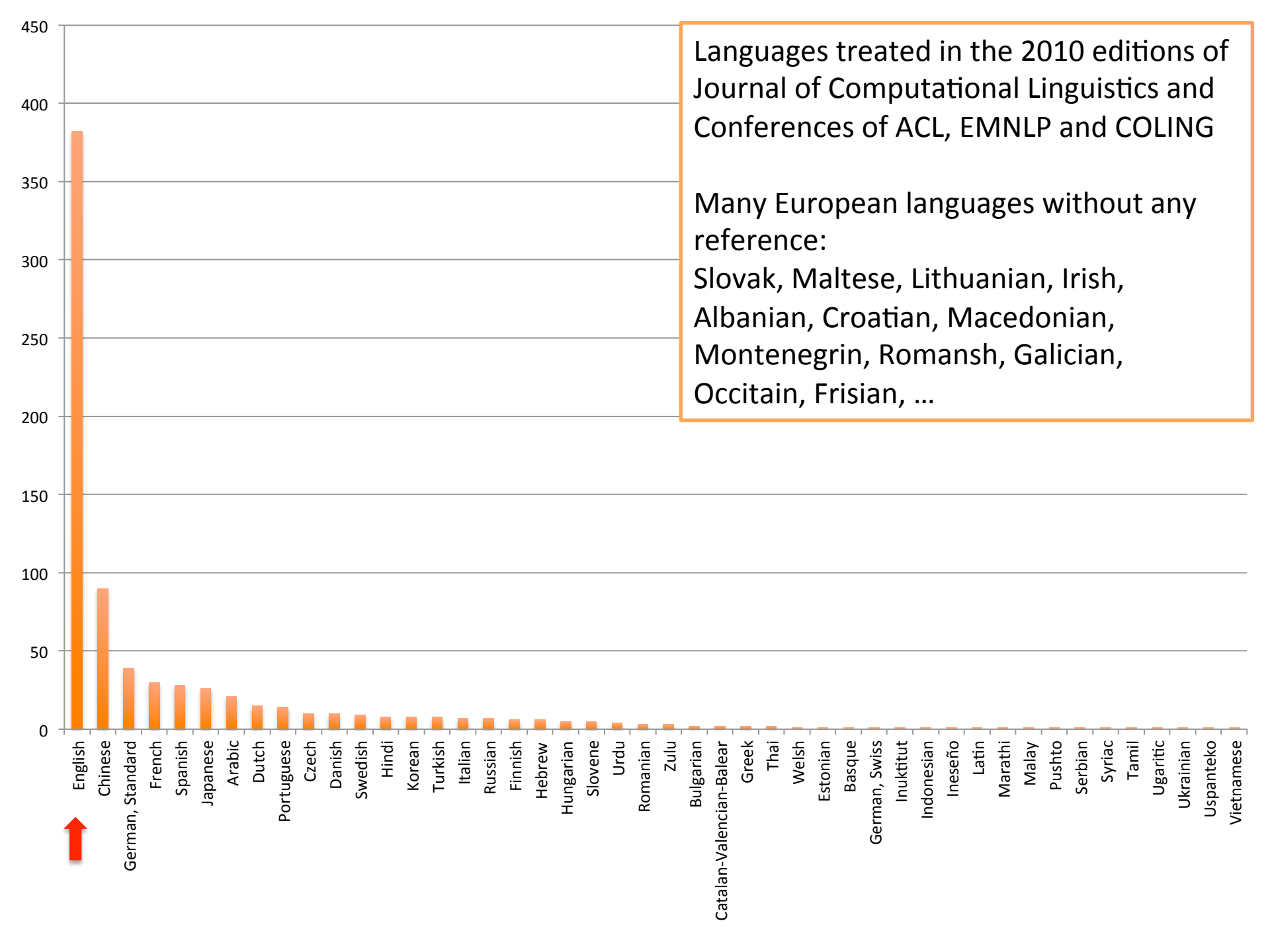
11: Text analysis: state of language technology support for 30 European languages

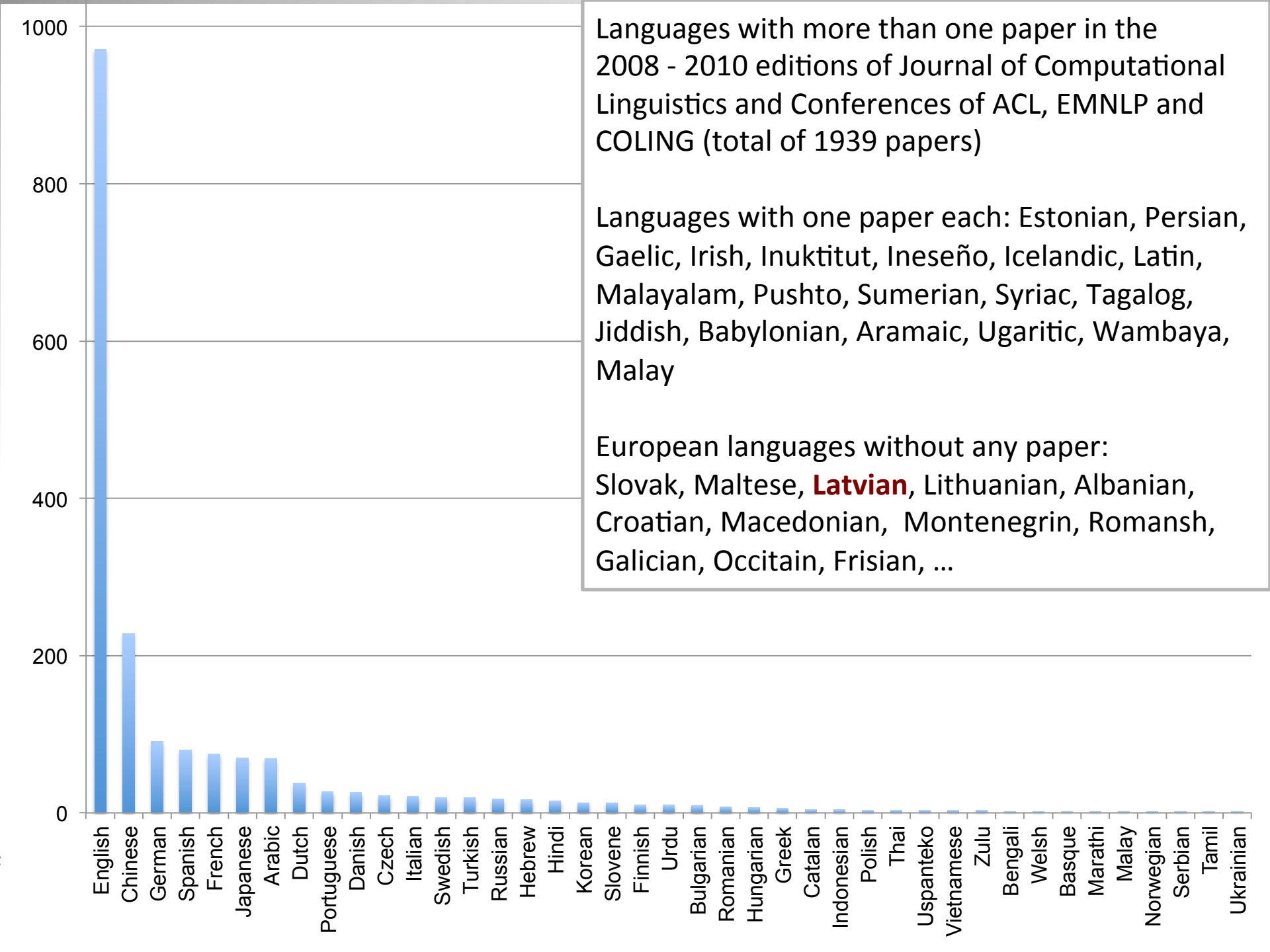
Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Czech Dutch Finnish French **German** Italian Portuguese Spanish	Basque Bulgarian Catalan Danish Estonian Galician Greek Hungarian Irish Norwegian Polish Serbian Slovak Slovene Swedish	Croatian Icelandic Latvian Lithuanian Maltese Romanian

9: Speech processing: state of language technology support for 30 European languages

Excellent support	Good support	Moderate support	Fragmentary support	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

12: Speech and text resources: State of support for 30 European languages


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, Macedonian, Montenegrin, Romansh, Galician, Occitain, Frisian, …
450
400
350
300
250
200
150
100
50
0
English
Chinese
German, Standard
French
Spanish
Japanese
Arabic
Dutch
Portuguese
Czech
Danish
Swedish
Hindi
Korean
Turkish
Italian
Russian
Finnish
Hebrew
Hungarian
Slovene
Urdu
Romanian
Zulu
Bulgarian
Catalan-Valencian-Balear
Greek
Thai
Welsh
Estonian
Basque
German, Swiss
Inuktitut
Indonesian
Ineseño
Latin
Marathi
Malay
Pushto
Serbian
Syriac
Tamil
Ugaritic
Ukrainian
Uspanteko
Vietnamese



Languages with more than one paper in the 2008 - 2010 editions of Journal of Computational Linguistics and Conferences of ACL, EMNLP and COLING (total of 1939 papers)

Languages with one paper each: Estonian, Persian, Gaelic, Irish, Inuktitut, Ineseño, Icelandic, Latin, Malayalam, Pushto, Sumerian, Syriac, Tagalog, Jiddish, Babylonian, Aramaic, Ugaritic, Wambaya, Malay

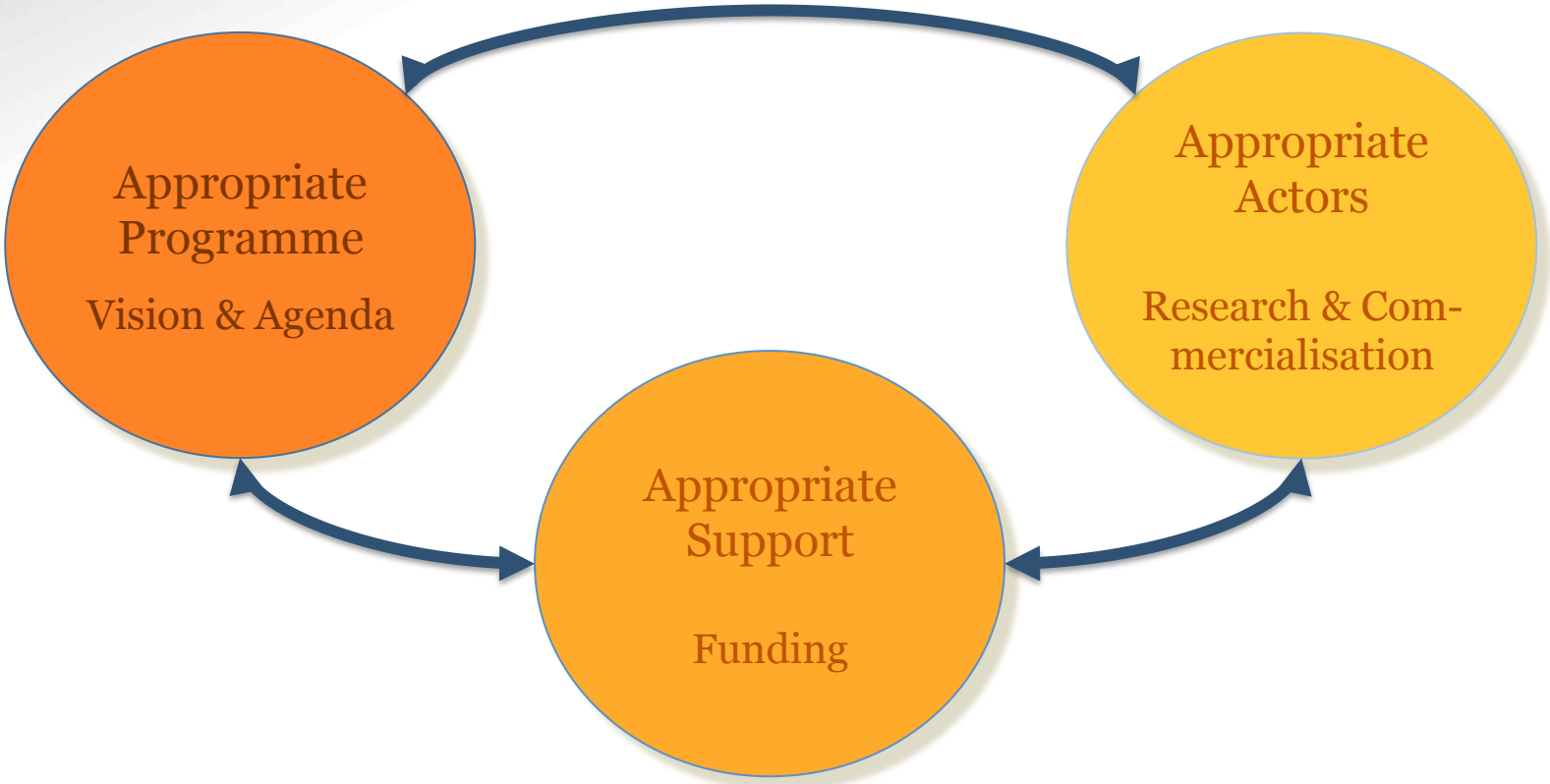
European languages without any paper:
Slovak, Maltese, **Latvian**, Lithuanian, Albanian, Croatian, Macedonian, Montenegrin, Romansh, Galician, Occitain, Frisian, …

Pitfalls

- Too much research in Europe follows patterns set by US research instead of concentrating on our own, demands, strengths and opportunities

- Example: Trying to follow DARPA Research and Google Translate instead of concentrating on European demands and strengths

We need a clear focussed program, a well coordinated community and adequate funding.



Three Lines of Action

META Multilingual Europe Technology Alliance

language days

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

Important steps have been taken

META Multilingual Europe Technology Alliance

- A Network of Excellence with 60 research centers in 34 countries
- An alliance, META, with 638 members (organizations) in 47 countries
- Vision Process with vision groups discussion at numerous conferences
- 30 Language White Papers on individual languages
- A first version of META-SHARE, the infrastructure for sharing resources
- A Strategic Research Agenda has been developed
- Inclusion of language communities - language policy bodies
- Inclusion of industrial and professional associations

A Platform for Resource Sharing

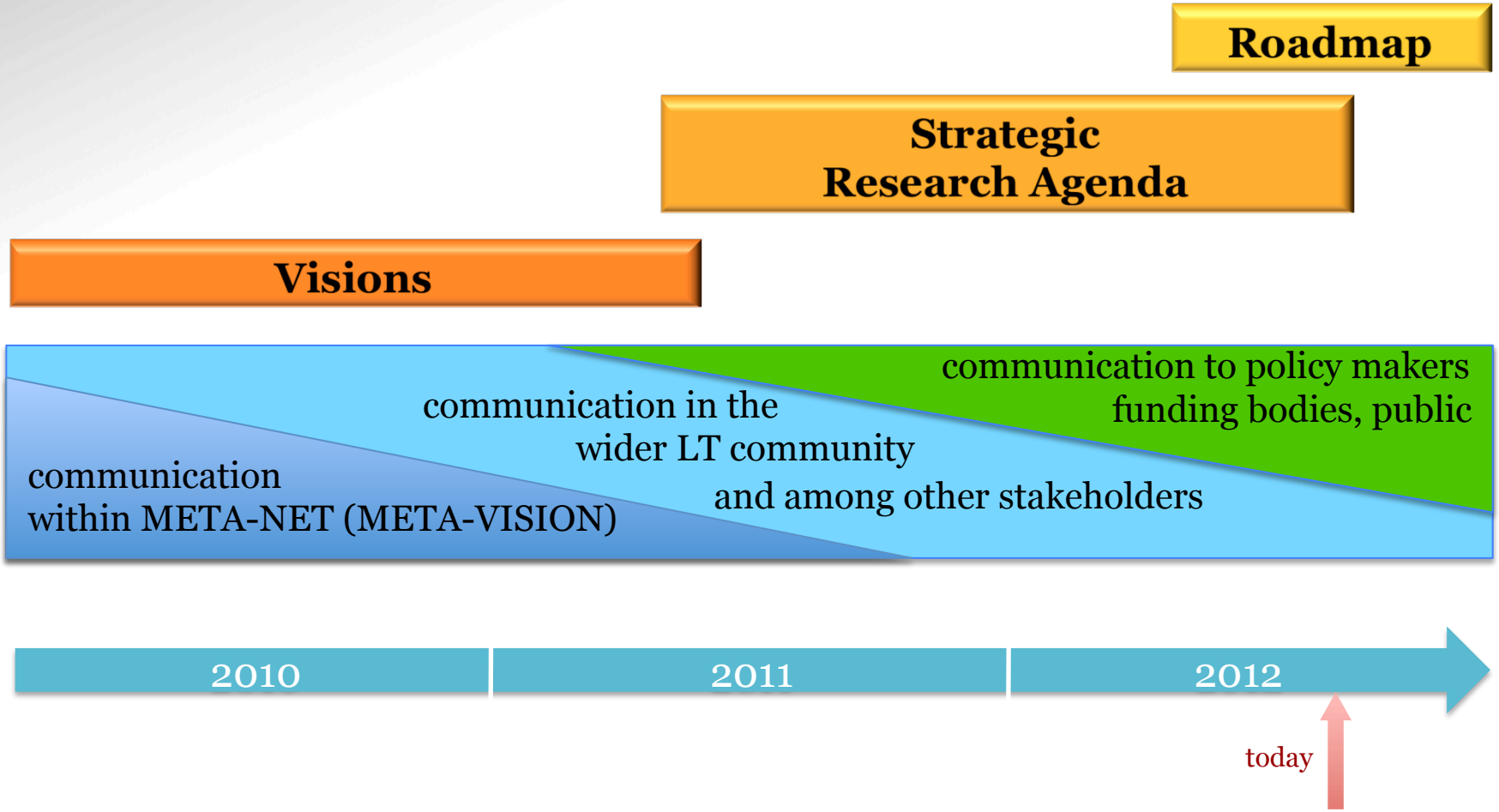
META Multilingual Europe Technology Alliance

language days

The Procedure

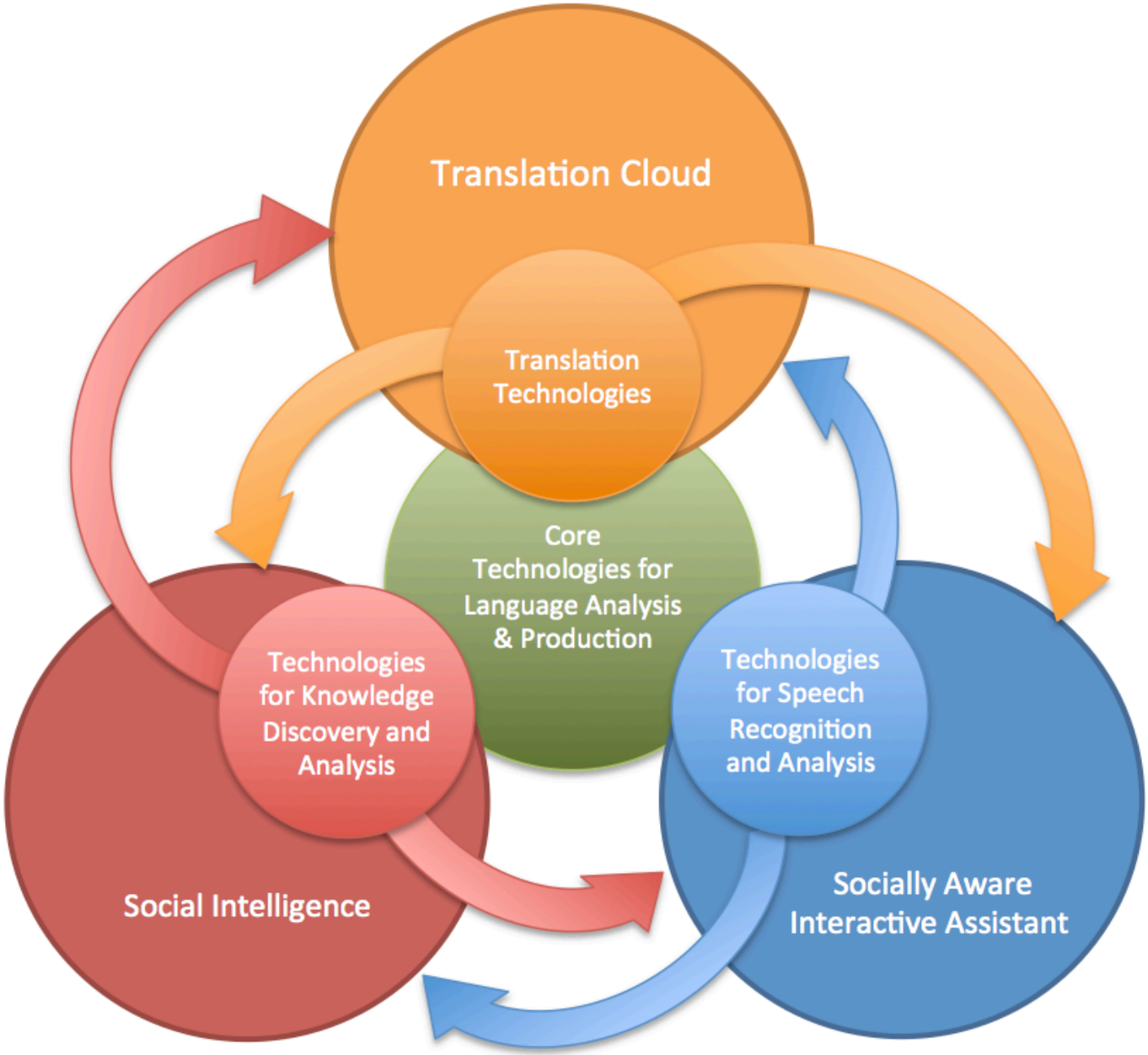
META Multilingual Europe Technology Alliance

language days

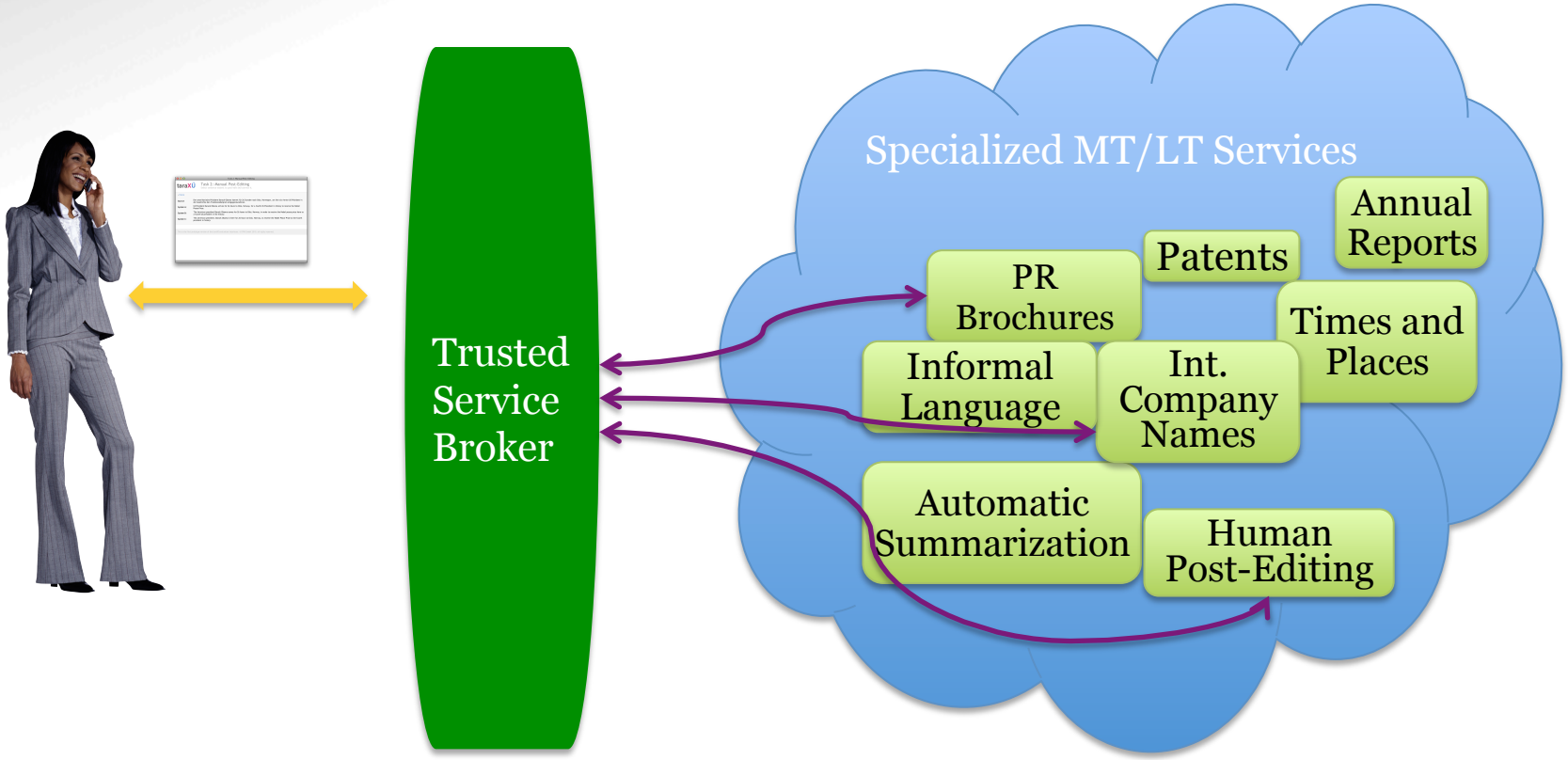


Priority Themes

- Translation Cloud – Understanding everything, everywhere, everytime
- Social Intelligence – Technologies for e-participation
- Second Me – Socially aware interactive assistant


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

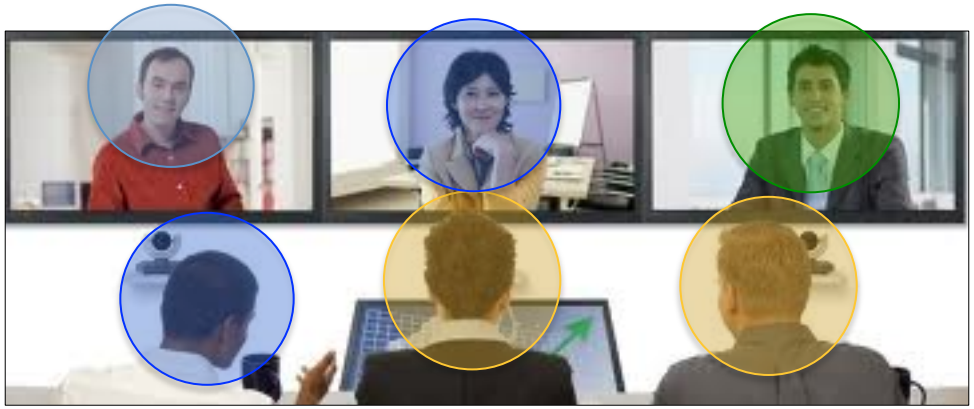
Translation Brokering



Ambient Translation Projection

Cross-lingual Virtual Meetings

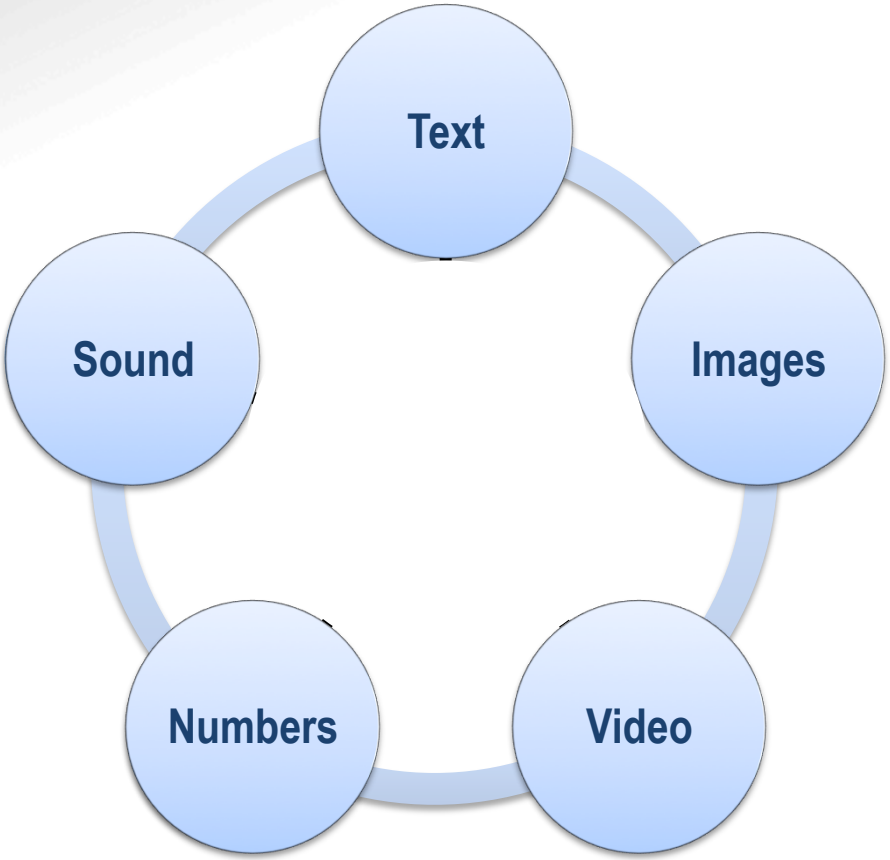
- Individual realtime translation of speech, slides, and handwritten text (shared whiteboard)
- Automatic minutes
- Searchable recordings
- Use cases:
  - Corporate
  - E-democracy
  - NGOs
  - Expert discussions
  - Fan clubs
  - Consumer fora
  - Medical self-help groups, etc.



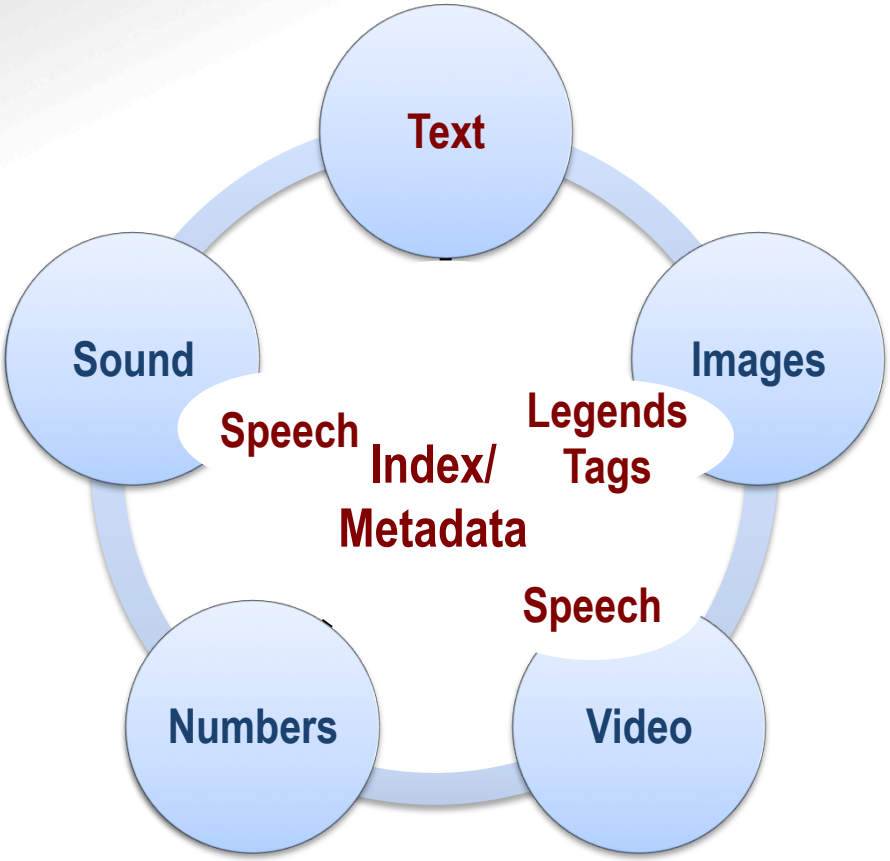
The Data View

- Data are the raw material from which information and knowledge are derived
- and thus analyses, decisions, strategies, solutions, policies, …
- There is a chain of processes that add value to the raw material

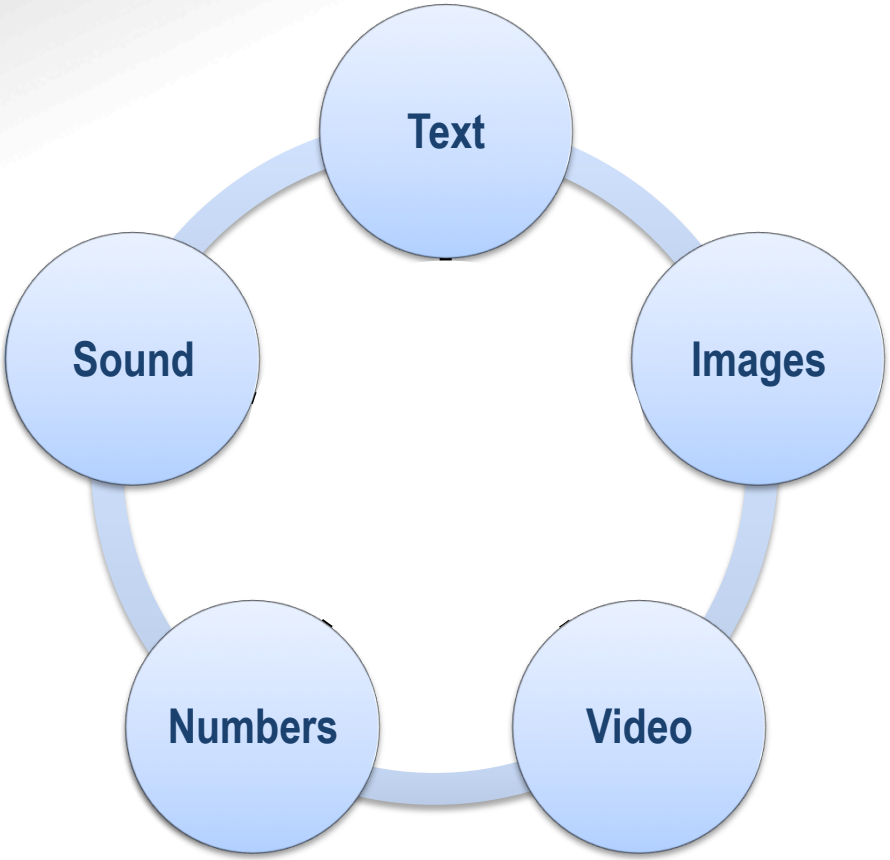
Language among other Data



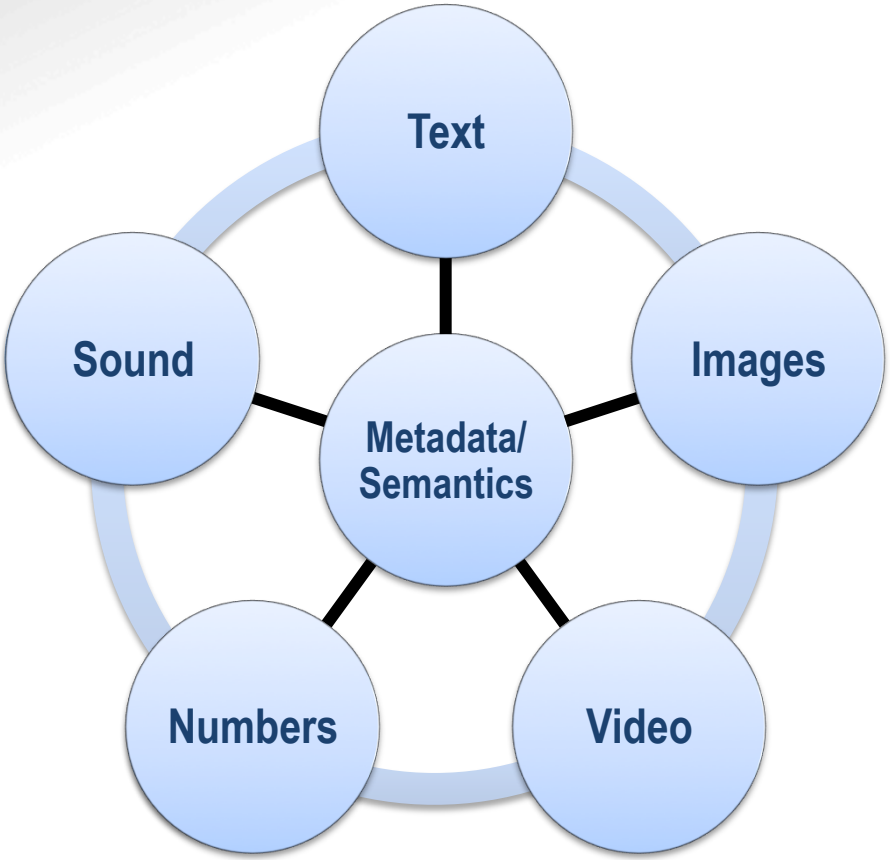
Language among other Data



Language among other Data



Language among other Data

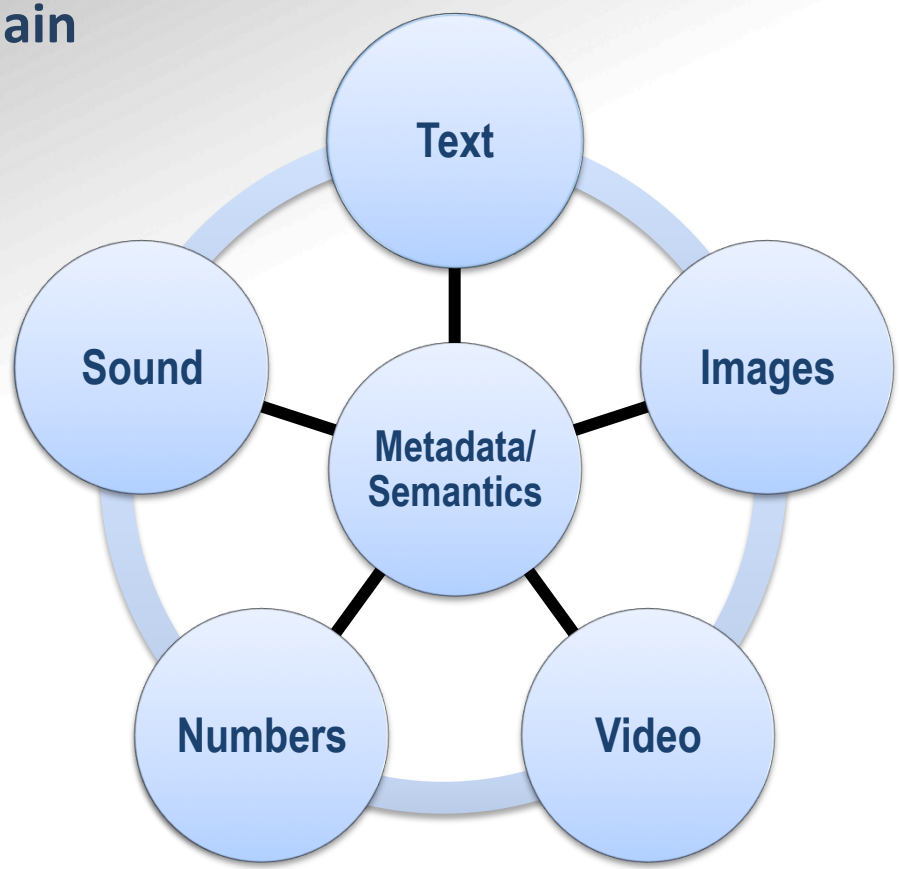


Metadata / Ontologies live on human language

<! - ერთი მიზანი ამ მაგალითად ontologies იყო, რომ ისინი იყოს იმი ქვეკლასით. ჩვენ მივიღეთ ღვინისა და საკვები ორმხრივად იმპორტი ერთმანეთთან, რადგან საკვები აქციების ბევრი ღვინის თვისებები. იმის ნაცვლად, რომ, ჩვენ შეიძლება არ გამოიყენება შემდეგი ასლები ყველა გაზიარებული ცნებები:

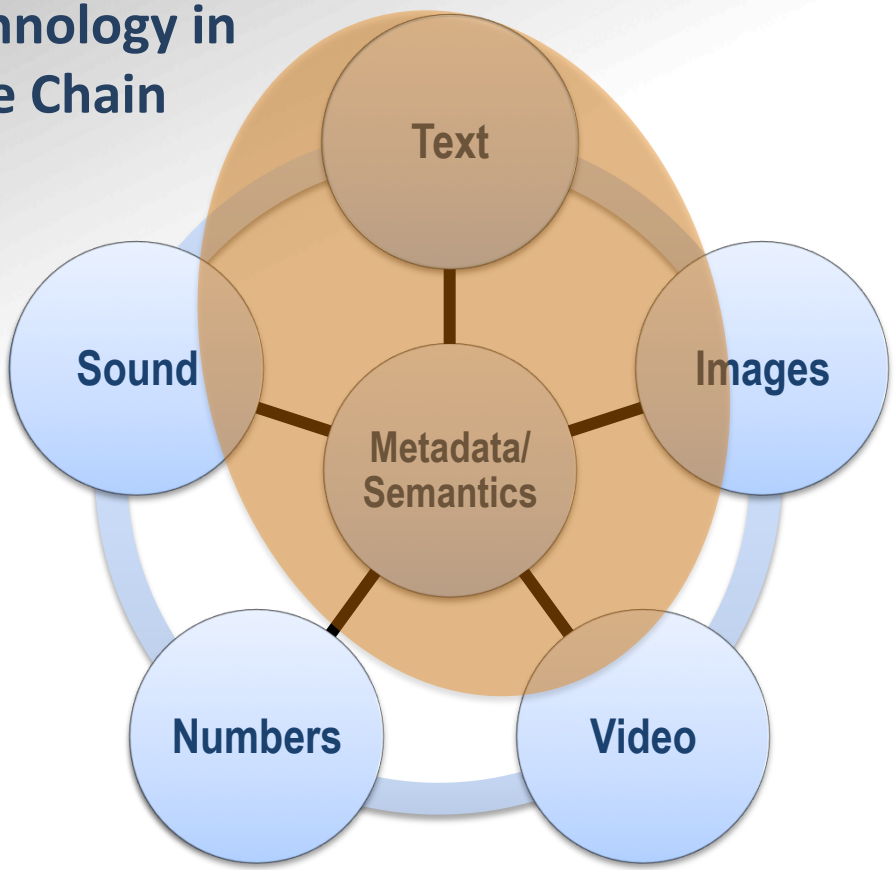
<ბუ:კლასი რაფ:პი="მაკარონი">
<ბუს:ქვეკლასით რაფ:რესურსი="#საკვებირამ" />
<ბუ:განსხვავდება რაფ:რესურსი="#ხორცი" />
<ბუ:განსხვავდება რაფ:რესურსი="#დოკუმენტისნახვა" />
<ბუ:განსხვავდება რაფ:რესურსი="#ზღვის" />
<ბუ:განსხვავდება რაფ:რესურსი="#დესერტი" />
<ბუ:განსხვავდება რაფ:რესურსი="#ხილის" />
</ ბუ: კლასი>

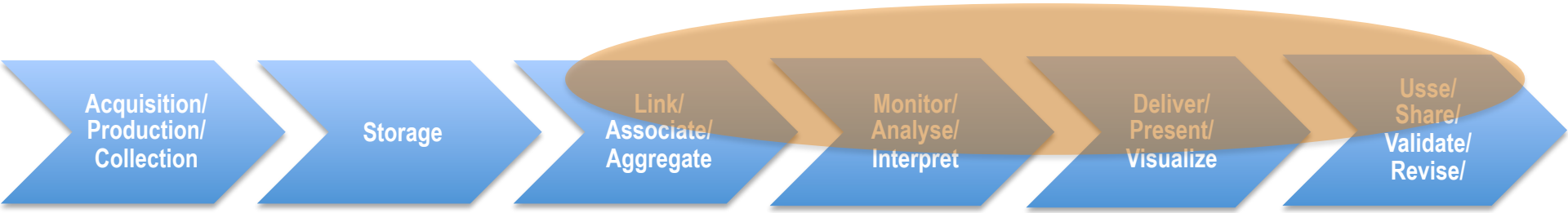
Data Value Chain



Acquisition/ Production/ Collection → Storage → Link/ Associate/ Aggregate → Monitor/ Analyse/ Interpret → Deliver/ Present/ Visualize → Usse/ Share/ Validate/ Revise/

Language Technology in the Data Value Chain





Social Intelligence & E-Participation

- Organizations need to understand their Constituents, i.e., their Cs: clients, consumers, citizens, customers, …
- To this end they need to interpret their opinions, communication, findings experience, demands, claims, …
- The Cs on the other hand, will get more power by participating in decision processes in politics, education, health care, markets
- They can take an active role: to this end they need to be informed on the issues and arguments, on newest developments, on the opinions and findings of others
- Functionalities:
  - Social intelligence by detecting and monitoring opinions, demands and needs
  - decision support for both decision makers and participants
  - Support of collective deliberation and collective knowledge accumulation

Our Next steps…

META Multilingual Europe Technology Alliance

- Present the Strategic Research Agenda
- Meet with national and EU research planners, funders and policy makers
- Address the public in as many member states as possible
- Mobilize user industries and administrations
- Realize the plan through research, innovation infrastructure programmes

What does ist cost?

- Such an effort does not come for free.
- It could easily cost as much as 100-150 km motorway.
- But no **additional** finances are needed !!!
- HORIZON 2020 and CEF could easily provide sufficient resources

  in H 2020:
  Inclusive, innovative and secure societies
  3.7 – 3.8 bEUR

Conclusion

- We have worked out a Strategic Research Agenda for Language Technology research and innovation
- that can put Europe ahead of its competitors in this important technology area and
- that will provide useful and attractive solutions to European society at the same time creating huge business opportunities for European industry

A Final Word

- My dentist jokingly warns:

  ”Save time:
  Only brush the teeth you want to keep.”

- This also holds true for language technology research and language support:

  “Save money:
  Only develop technologies for languages you really want to keep alive.”


META
Liels paldies!