META Multilingual Europe Technology Alliance

Language Technology for Europe: Chances, Needs and Plans

Hans Uszkoreit

Coordinator META-NET

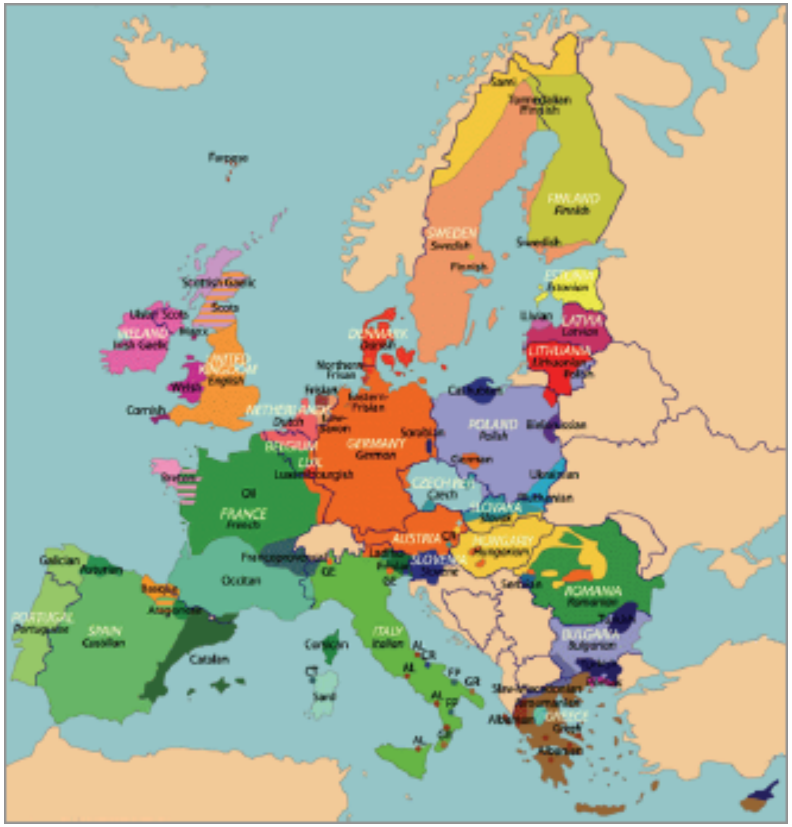
CESAR

Roadshow 2012 Sofia, May 2nd 2012

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 flow 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

The Role of Technology

- → 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

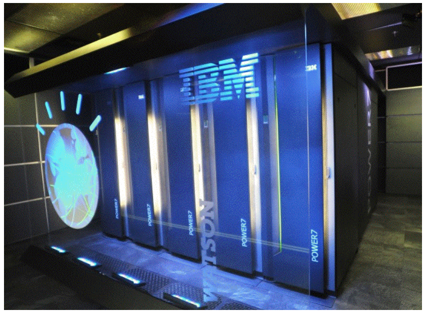
Today LT is already surrounding us

- → 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

Οι ισχύοντες κανόνες, ωστόσο, θεσπίστηκαν γ
οποία μάλιστα δεν είχε να αντιμετωπίσει παγκ

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



But this is only the beginning…

… 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

- → 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.

Future Internet and Multilingual Web

→ 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”)

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

- → 20 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…

… 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

- → 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

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

but it is fragmented

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

Europe has greater demand

- 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 Trados, Google Translate, LocalizationWorld, 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

→ LT research on European languages, except for English, is too weak and too slow.

→ Many languages are badly covered.

Findings of 30 Language White Papers

META Multilingual Europe Technology Alliance

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

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

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

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

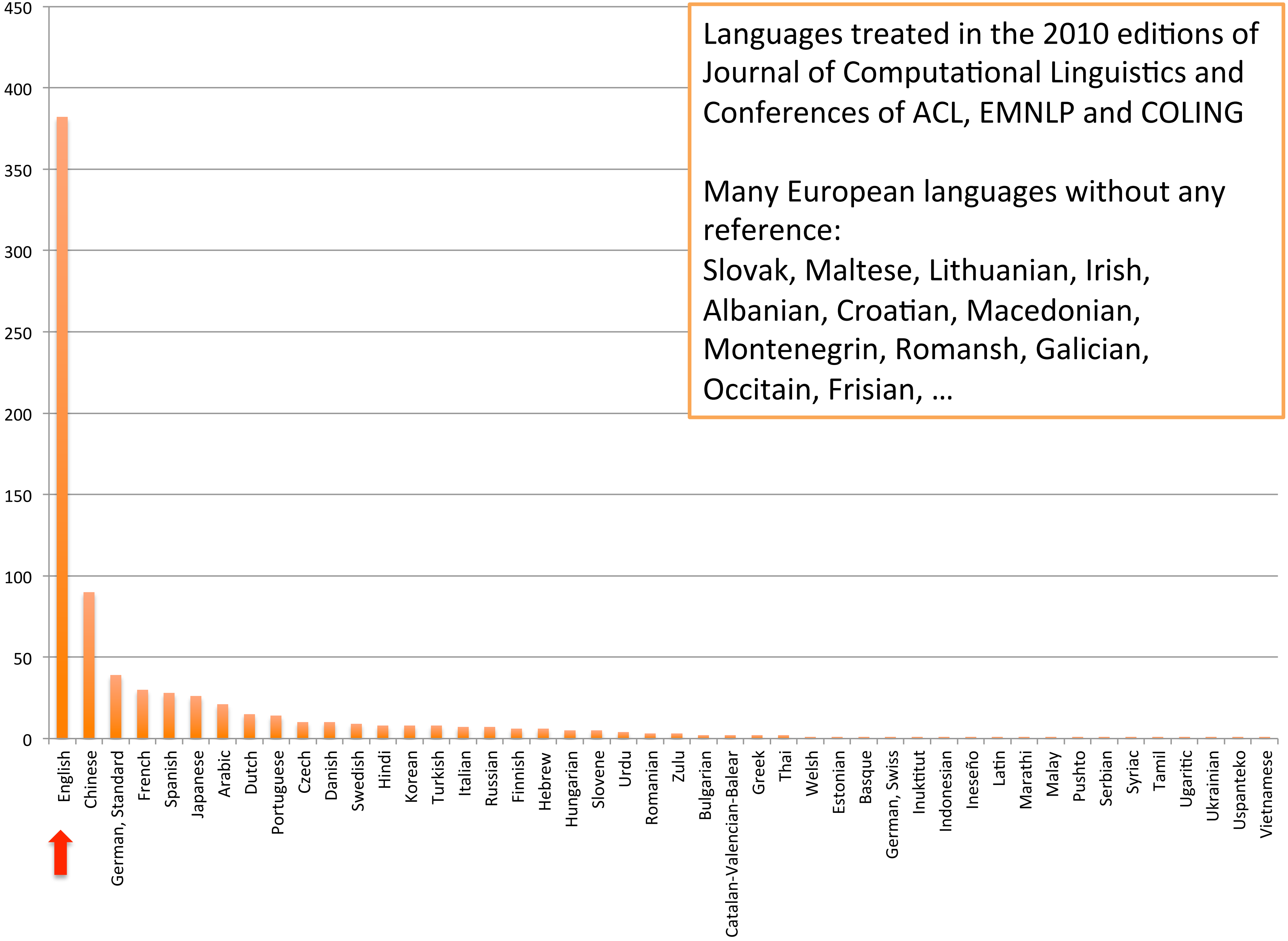
11: Text analysis: state of language technology support for 30 European languages

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

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

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

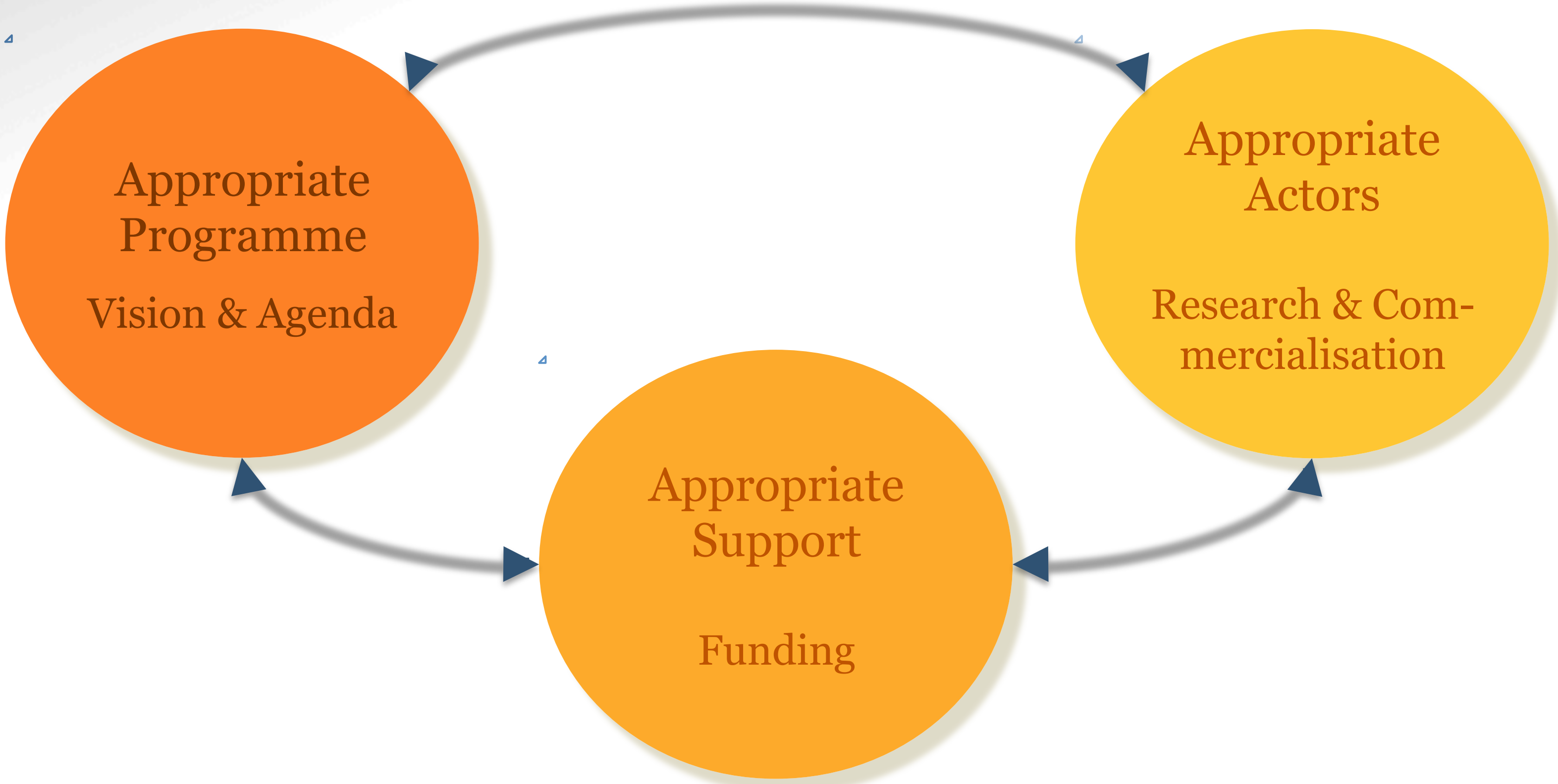
Many European languages without any reference:
Slovak, Maltese, Lithuanian, Irish, Albanian, Croatian, Macedonian, Montenegrin, Romansh, Galician, Occitain, Frisian, …

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.



Important steps have been taken

- → A Network of Excellence with 54 research centers in 33 countries
- → An alliance, META, with 307 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
- → Basic Outline of a Strategic Research Agenda
- → Inclusion of language communities - language policy bodies
- → Inclusion of professional associations

Three Lines of Action

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

META-SHARE: Building an open resource exchange infrastructure

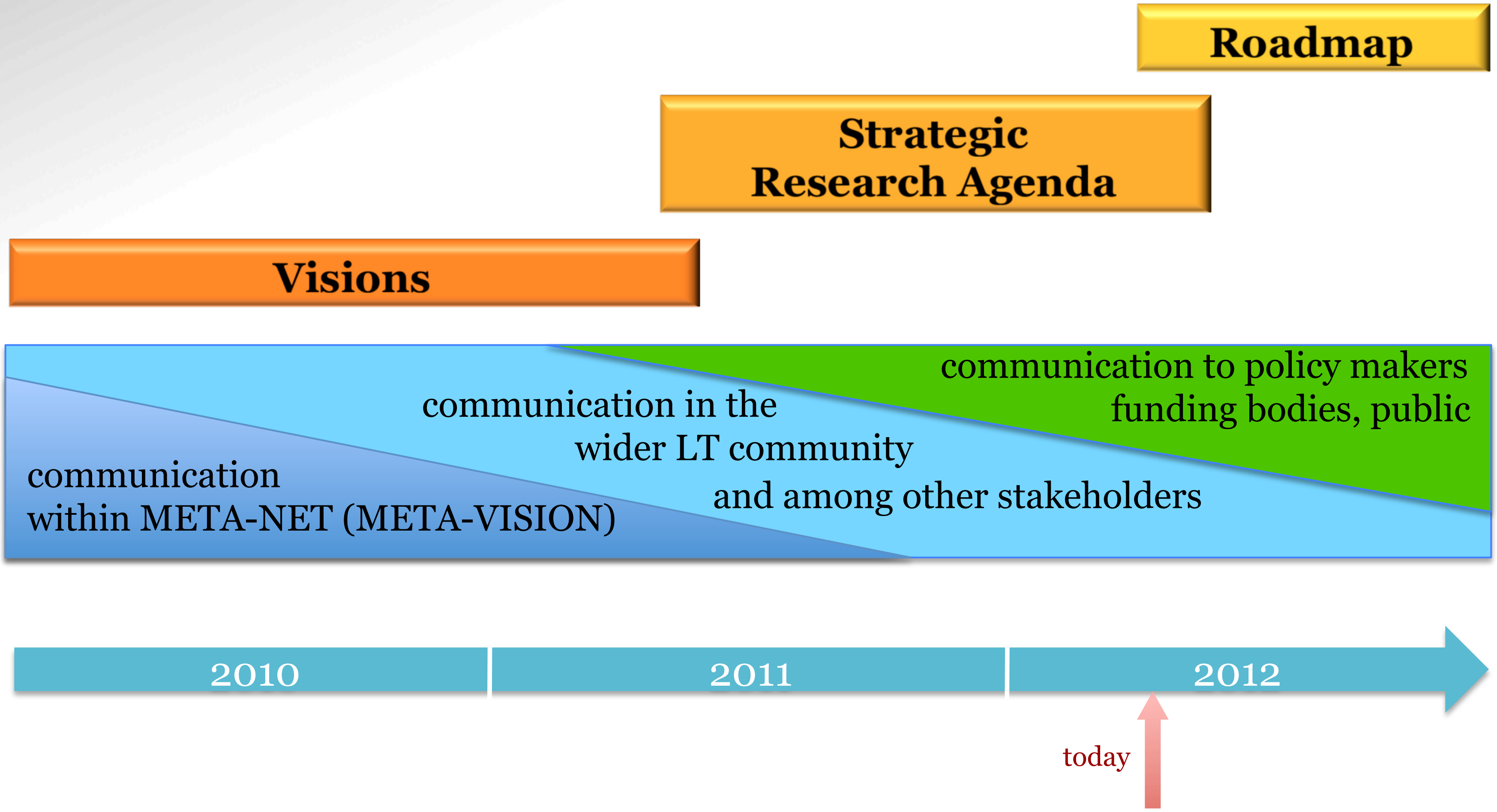
META-RESEARCH: Building bridges to neighbouring technology fields

A Platform for Resource Sharing

META SHARE

- → META-SHARE metadata specification version 1.0 published
- → META-SHARE prototype implemented and tested
- → Licensing policy and licensing templates
- → Charter for Language Resource Sharing

The Procedure



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

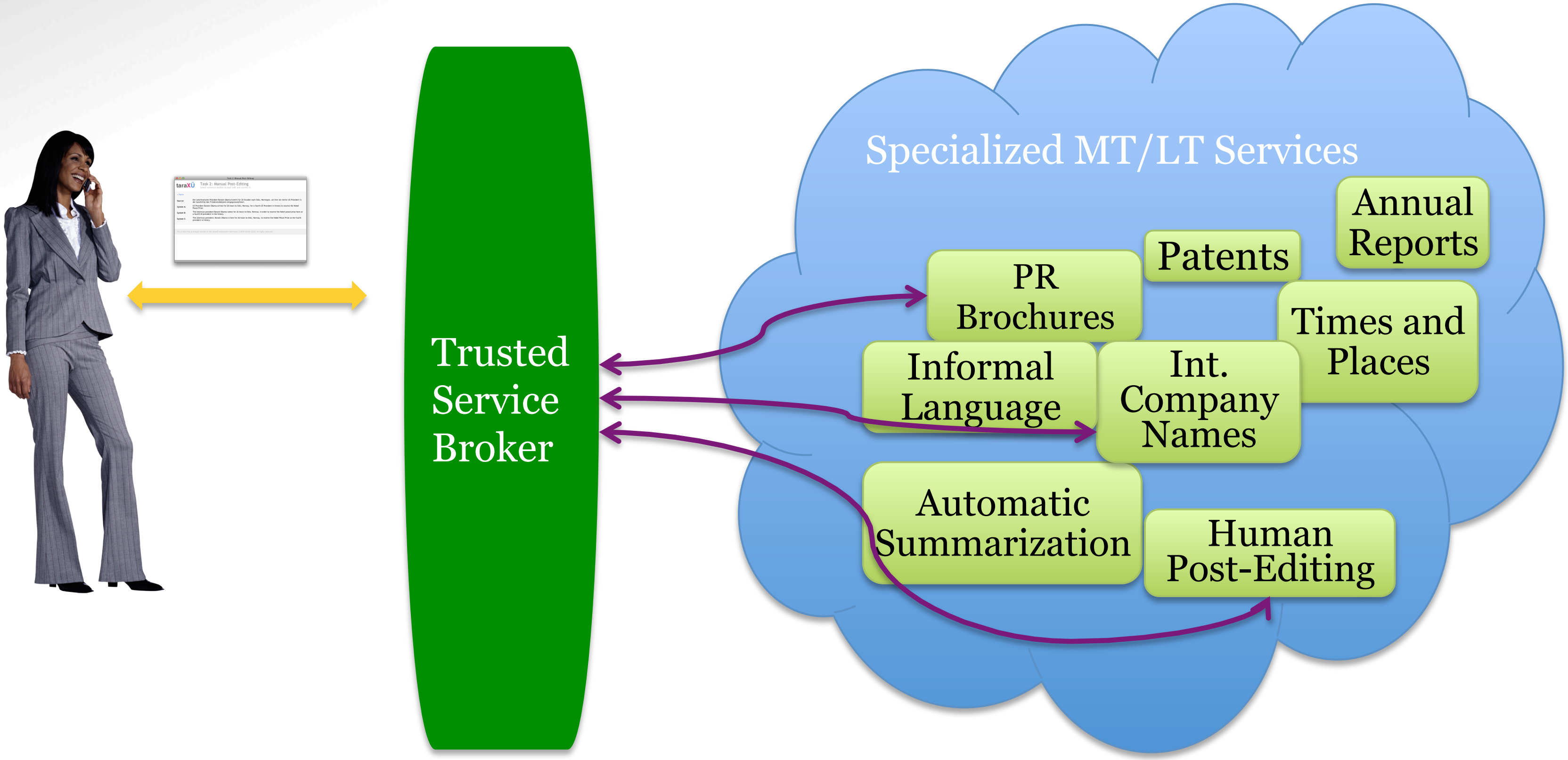
2012

today

Priority Themes

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

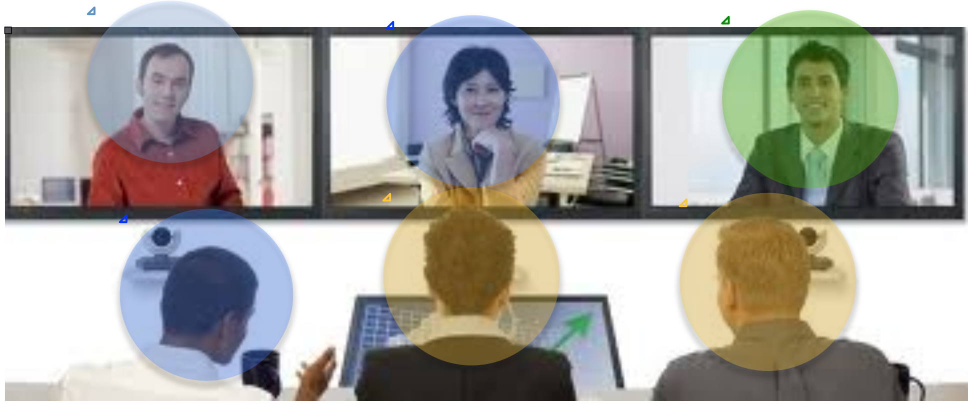
Translation Brokering



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.



Conclusion

- → We are confident that we will together with all stakeholders present in December a Strategic Research Program
- → 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

What does it cost?

- Because of progress in multilingual technologies and common core technologies, the funding needed for all the applications and all the languages does not propel us out of the range of other technology areas.
- Our collective estimates are 500-600 MEURO for R&I over the next six years
- 200-300 MEURO for Infrastructure and common services, also as a means to leverage national and industrial investment.

Next steps...

- → Finish the Strategic Research Agenda
- → Meet with national research planners, funders and policy makers
- → Address the public in as many member states as possible
- → Mobilize user industries and administrations

It should be possible because...

→ 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

→ But instead of reducing our focus to current hype themes such as **Big Data** and **Cloud Computing**

→ We should understand that **the most demanding and interesting data are are encoded in human languages**

→ And that one of the best examples of a **truly European Cloud solution is language services overcoming language boundaries**

To conclude:

→ For roughly the costs of 20-100 km motorway (depending on bridges and tunnels) …

→ … Europe could solve pressing problems and grab a unique opportunity

→ Get ahead of competitors in a technology of the near future that is not yet dominated by others

→ …that will be the key to human-centered IT

→ …and that will enable Europe to meet the constitutional obligations toward inclusion of all citizens, equality of cultures, and preservation of languages

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