META-NET Strategic Research Agenda

John Judge

CNGL, Dublin, Ireland

jjudge@computing.dcu.ie

Kites Symposium 2013
Helsinki, Finland – October 30, 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

- Introduction
- Language White Paper Series: Europe's Languages in the Digital Age
- The META-NET Strategic Research Agenda for Multilingual Europe
- Conclusions

Multilingual Europe

META-NET

- Where were we back in 2010?
- **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 in a strategic alliance for a major dedicated push**.

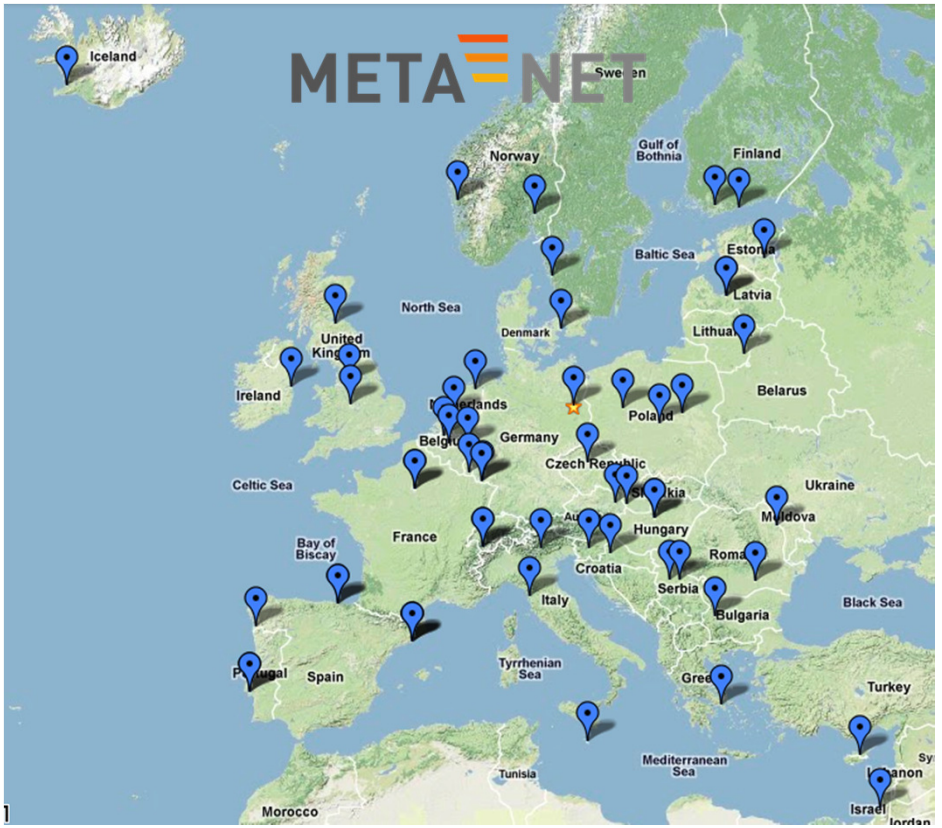
Objectives

META-NET

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

Four EU-Funded Projects

- Initial project: T4ME (FP7; 13 partners, 10 countries)
- Three ICT-PSP consortia since Feb. 2011: CESAR, METANET4U, META-NORD
- All four projects ended on January 31, 2013.
- All EU member states and several non-member states covered.
- META-NET in Sept. 2013: **60** members in **34** countries.



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

META NET

Europe's Languages in the Digital Age

Language White Paper Series

Language White Paper Series

- "Europe's Languages in the Digital Age".
- Reports on the state of our languages in the digital age and the level of support through language technology.
- Series covers 32 languages.
- Key communication instruments to address decision makers and journalists.
- Inform about societal and technological problems and challenges as well as economic opportunities.
- >2 years in the making.
- >200 national experts as contributors.
- >8000 copies printed and distributed to politicians and journalists.

White Paper Box Sets (100 copies)

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

Just Added: Welsh

* = Official EU language

Cross-Lingual Comparison

- 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 Analytics
  4. Language Resources
- Results finalised at a meeting in Berlin with representatives of all 30 languages (October 21/22, 2011).

MT

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

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

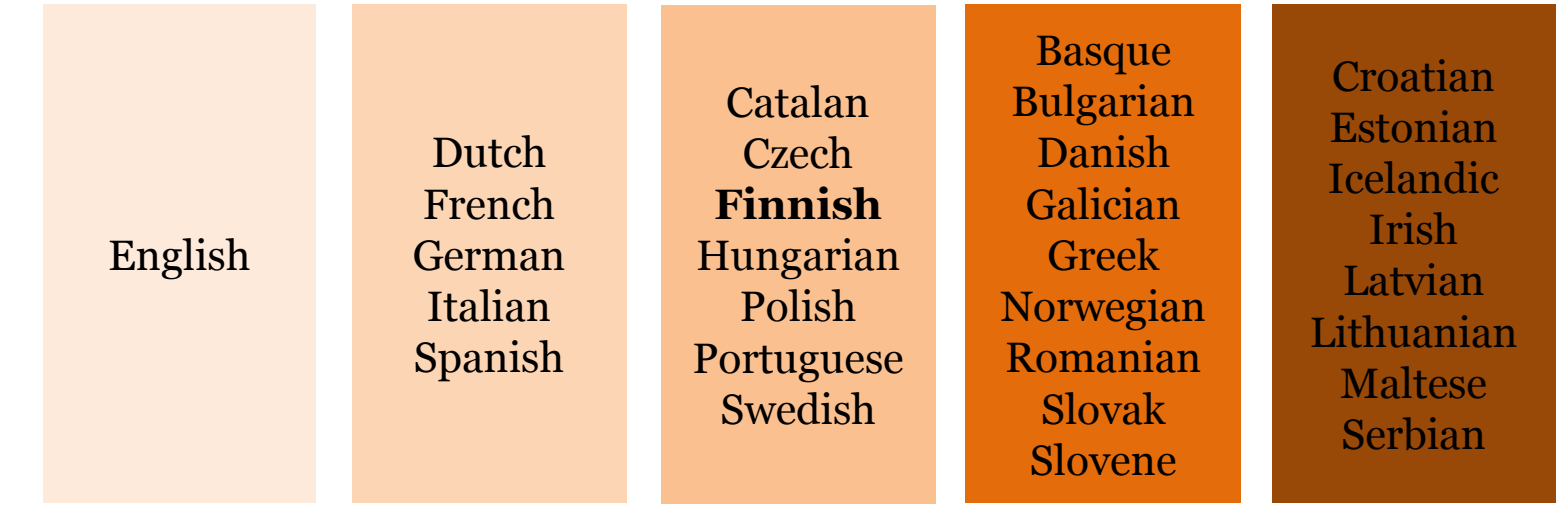
Speech

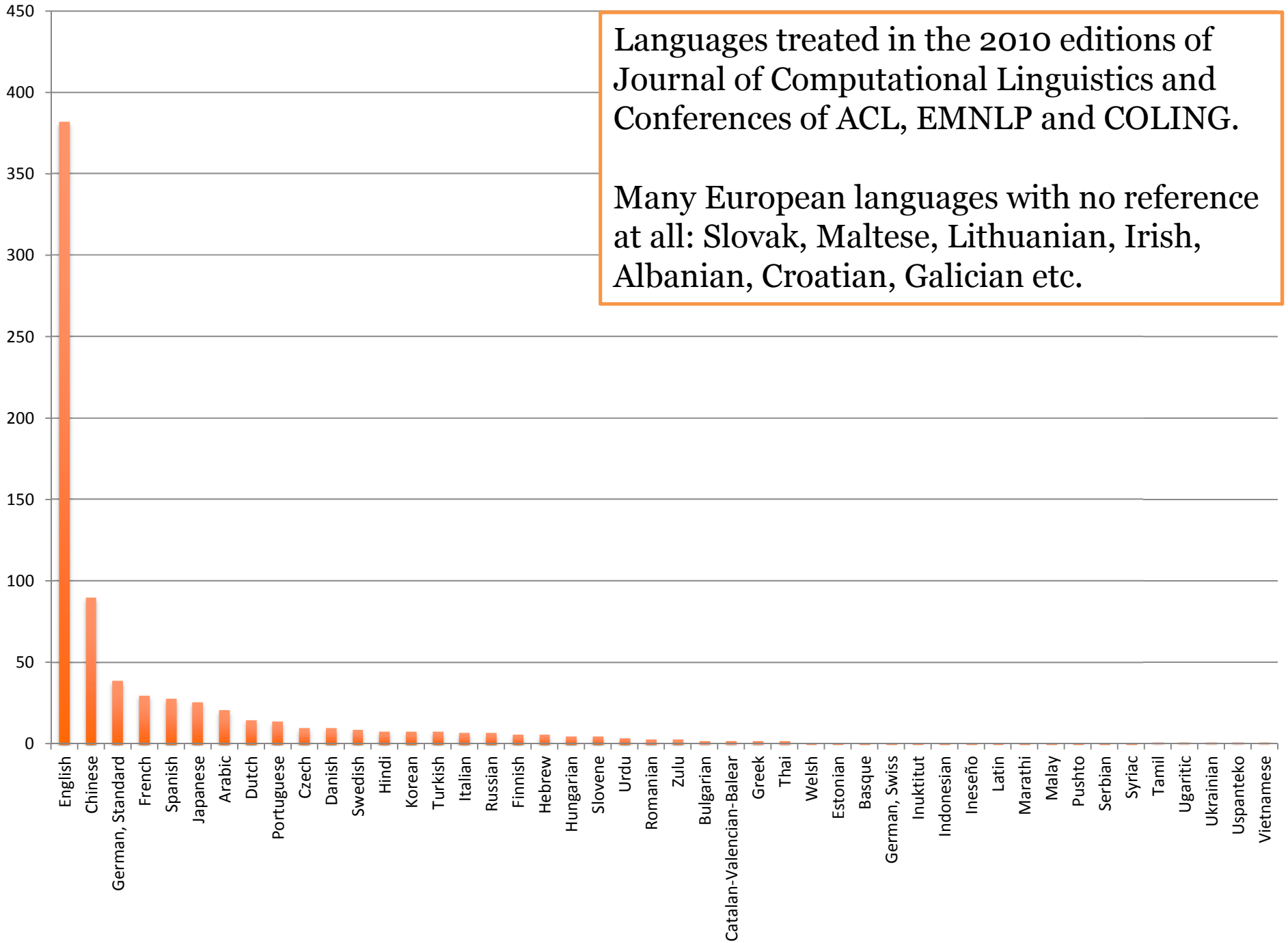
excellent	good	moderate	fragmentary	weak or no support
	English	Czech, Dutch, **Finnish**, French, German, Italian, Portuguese, Spanish	Basque, Bulgarian, Catalan, Danish, Estonian, Galician, Greek, Hungarian, Irish, Norwegian, Polish, Serbian, Slovak, Slovene, Swedish	Croatian, Icelandic, Latvian, Lithuanian, 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

Europe's Languages and LT





Languages treated in the 2010 editions of Journal of Computational Linguistics and Conferences of ACL, EMNLP and COLING.

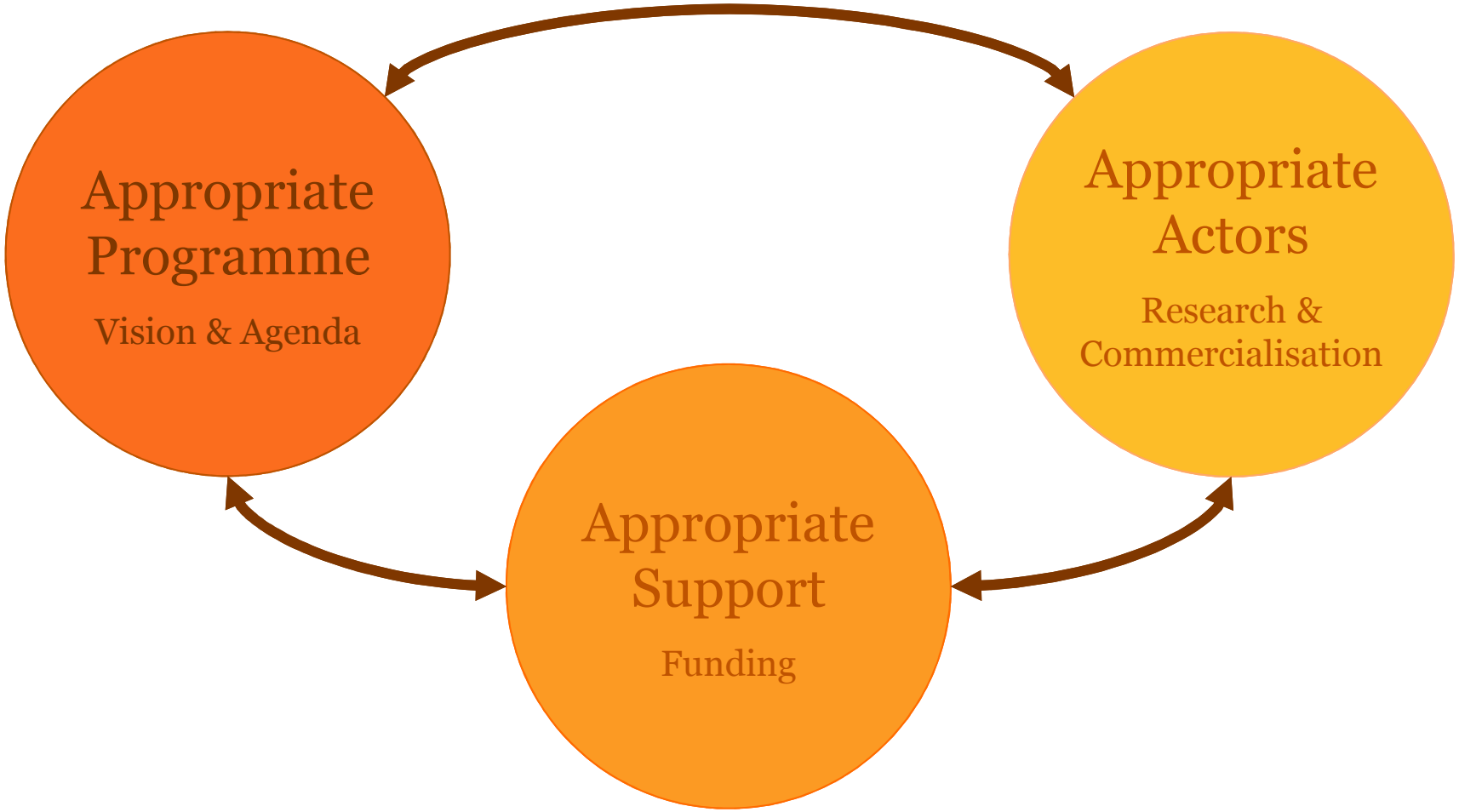
Many European languages with no reference at all: Slovak, Maltese, Lithuanian, Irish, Albanian, Croatian, Galician etc.

Key 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 – Icelandic, Latvian, Lithuanian, Maltese – receive this weakest score in *all four areas!*
- *At least 21 European languages in danger of digital extinction!*(Languages put into the "weak or no support" category at least once.)

Strategic Research Agenda

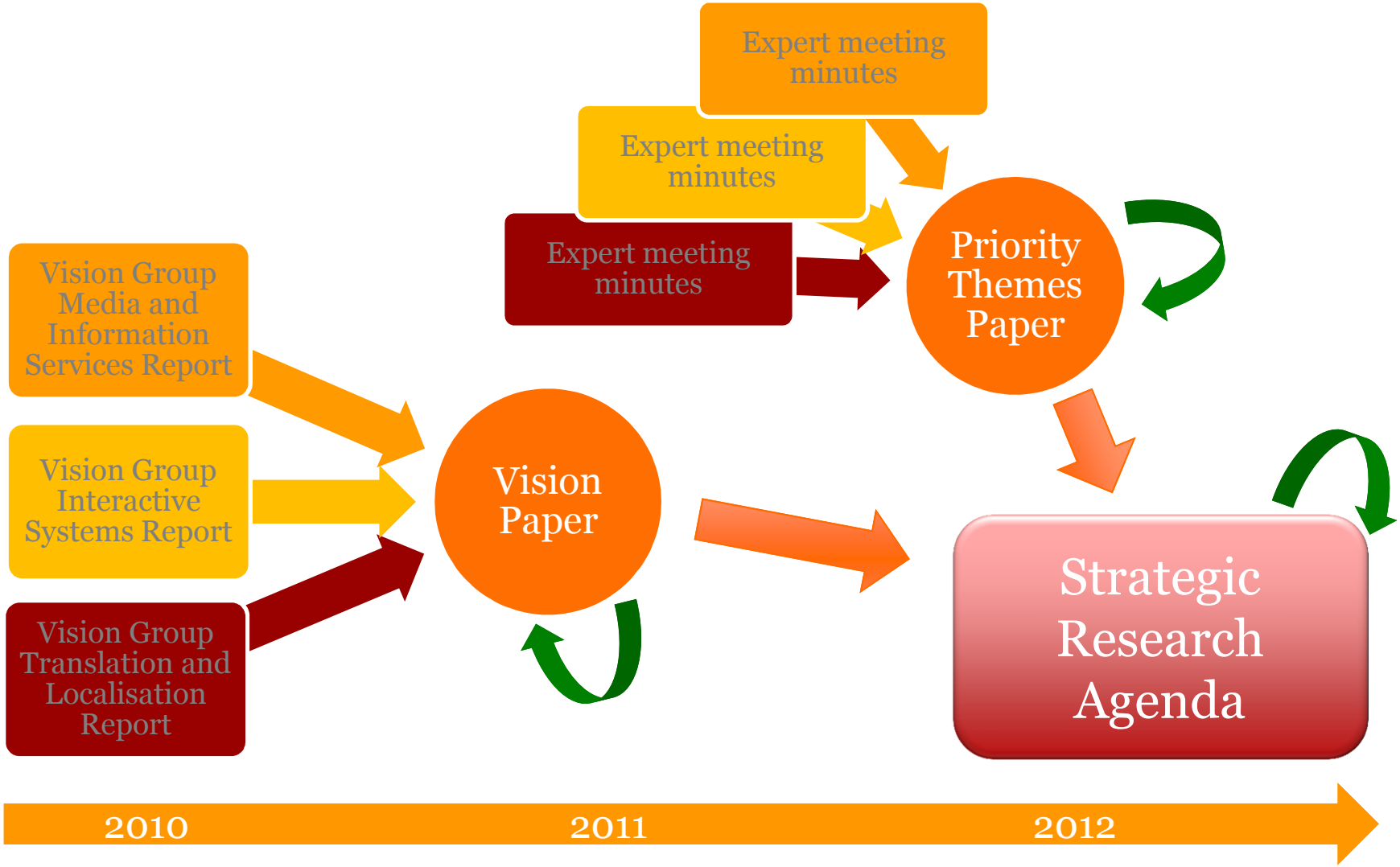
Three Ingredients



Three Vision Groups

- **Translation and Localisation** (technical documentation, official bulletins, GUI localisation, games, services etc.)
  - Target stakeholders: large users of translation services, (machine) translation, software companies, game companies, localisation industry
- **Media and Information Services** (audiovisual sector, news, digital libraries, portals, search engines etc.)
  - Target stakeholders: media industries, search engine providers, archives
- **Interactive Systems** (mobile assistance, dialogue translation, call centres, etc.)
  - Target stakeholders: mobile software and service providers, telecom industry, call centres

Planning Process



Preparation of the SRA

- Strategic Research Agendas of other initiatives were screened.
- Many suggestions as input from Vision Group members.
- We discussed procedures, input and structure of the SRA in four meetings of the META Technology Council.
  - Brussels, Belgium, November 16, 2010
  - Venice, Italy, May 25, 2011
  - Berlin, Germany, September 30, 2011
  - Brussels, Belgium, June 19, 2012
- Additional input in talks, meetings, workshops, discussions, etc.
  - Example: Three HLT Expert Meetings organised by the EC (end of 2011)
- Almost 200 experts contributed to the SRA (54% from industry; 46% from research; 4% from national/international institutions).

Strategic Research Agenda

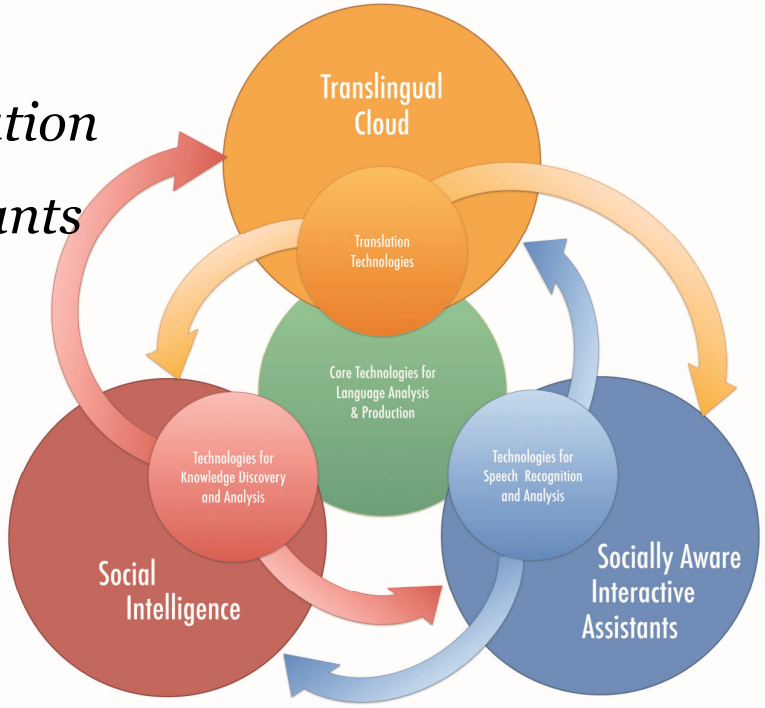
- Addresses the problems we identified when preparing the white papers.
- Three priority research themes and application/innovation scenarios.
- Can put Europe ahead of its competitors in this technology area.
- >190 contributors; >2 years.
- Presented and discussed at 83 conferences and major workshops.
- Final version ready on Dec. 1, 2012.
- http://www.meta-net.eu/sra

SRA: Contents – Brief Glimpse

- Set the stage and describe the European situation, the needs and the LT research and industry.
- Discuss the state of IT, predictions and mega-trends.
- Our technology vision for 2020.
- Select and specify priority themes.
- Suggest a model for speeding up innovation.
- Outline proposals for the organisation of research and innovation.

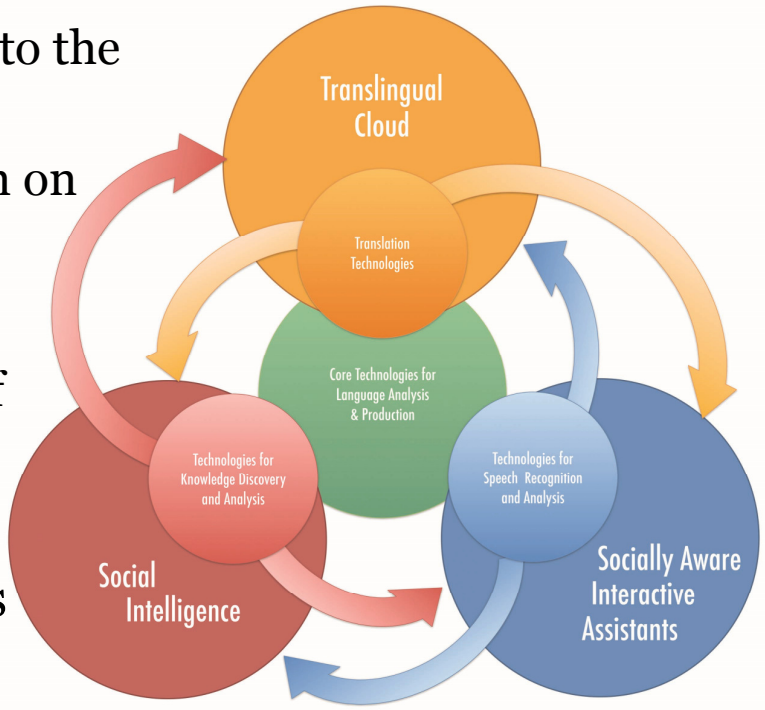
Priority Themes: 3 + 2

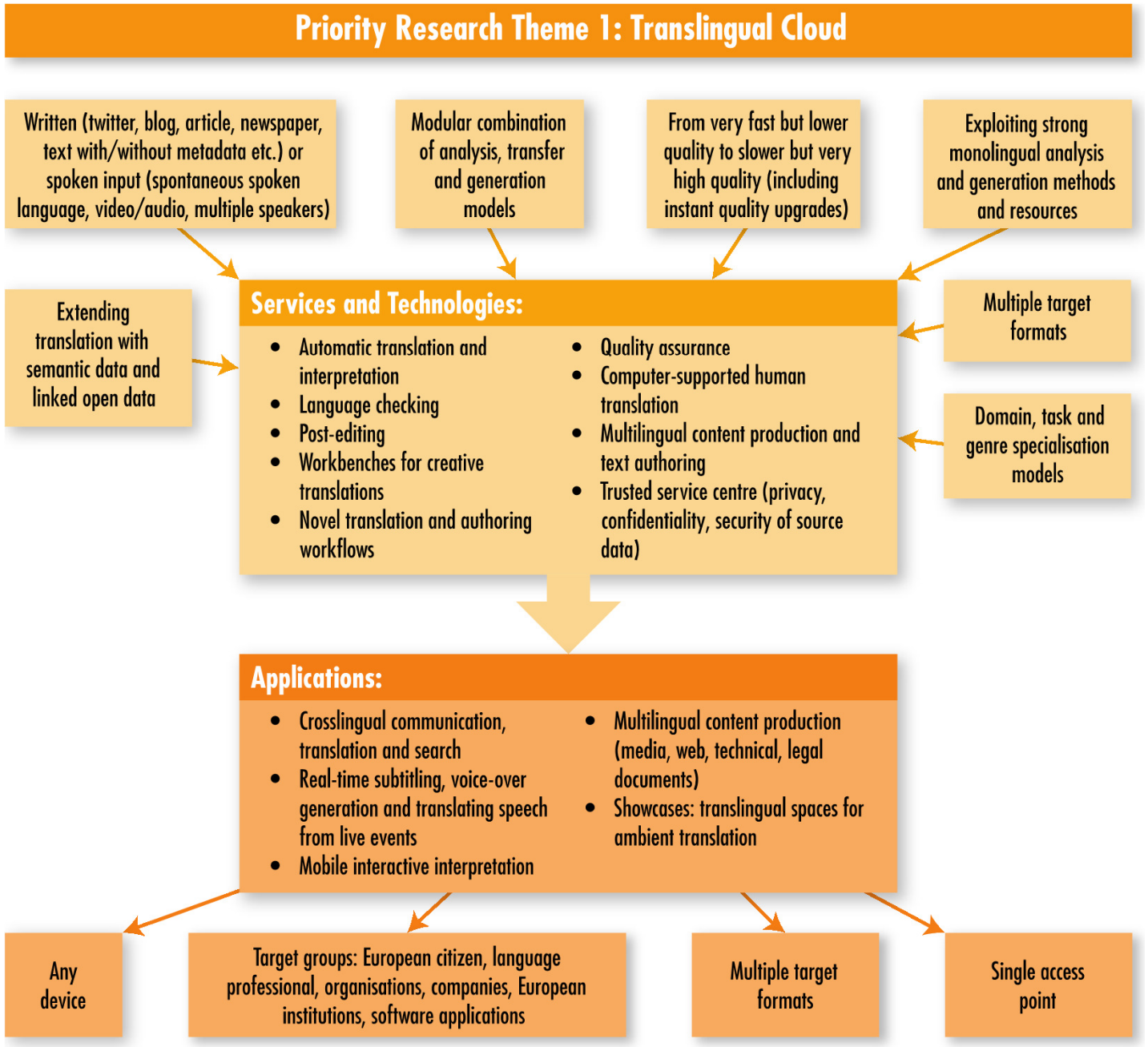
- We decided on priority themes that (a) support technology progress, (b) lead to solutions that European society needs and (c) solutions from which European industry will benefit as users or as providers.
  - *Translingual Cloud*
  - *Social Intelligence and e-Participation*
  - *Socially-Aware Interactive Assistants*
- Two additional themes:
  - *European Service Platform for Language Technologies*
  - *Core Technologies for Language Analysis and Production*



PT1: Translingual Cloud

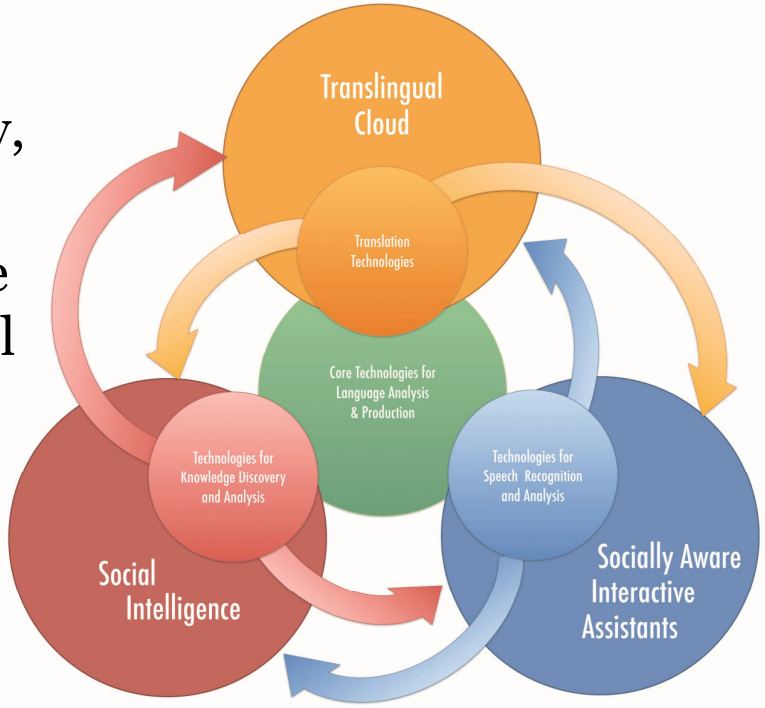
- Europe has a big need for translations of publishable quality.
- Focus on high-quality translation.
- New research paradigms
  - Inclusion of professional translators into the research process
  - Inclusion of technologists into research on human translation processes
- Different technological approaches
  - Stronger emphasis on the properties of individual languages
  - A central role for semantics
- Methods for specific genres & domains

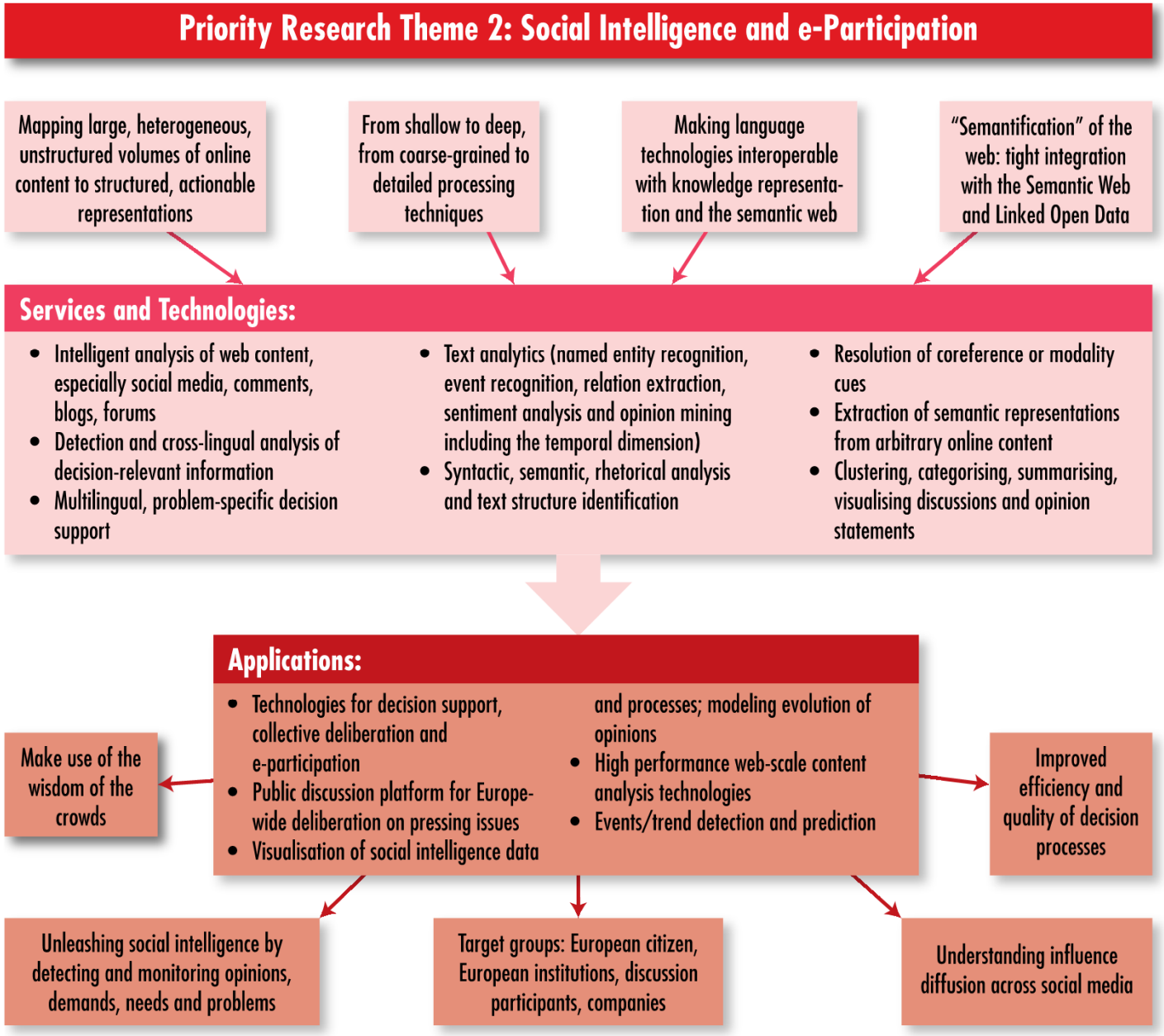



Priority Research Theme 1: Translingual Cloud
Written (twitter, blog, article, newspaper, text with/without metadata etc.) or spoken input (spontaneous spoken language, video/audio, multiple speakers)
Modular combination of analysis, transfer and generation models
From very fast but lower quality to slower but very high quality (including instant quality upgrades)
Exploiting strong monolingual analysis and generation methods and resources
Extending translation with semantic data and linked open data
Services and Technologies:
Automatic translation and interpretation
Language checking
Post-editing
Workbenches for creative translations
Novel translation and authoring workflows
Quality assurance
Computer-supported human translation
Multilingual content production and text authoring
Trusted service centre (privacy, confidentiality, security of source data)
Multiple target formats
Domain, task and genre specialisation models
Applications:
Crosslingual communication, translation and search
Real-time subtitling, voice-over generation and translating speech from live events
Mobile interactive interpretation
Multilingual content production (media, web, technical, legal documents)
Showcases: translingual spaces for ambient translation
Any device
Target groups: European citizen, language professional, organisations, companies, European institutions, software applications
Multiple target formats
Single access point

PT2: Social Intelligence

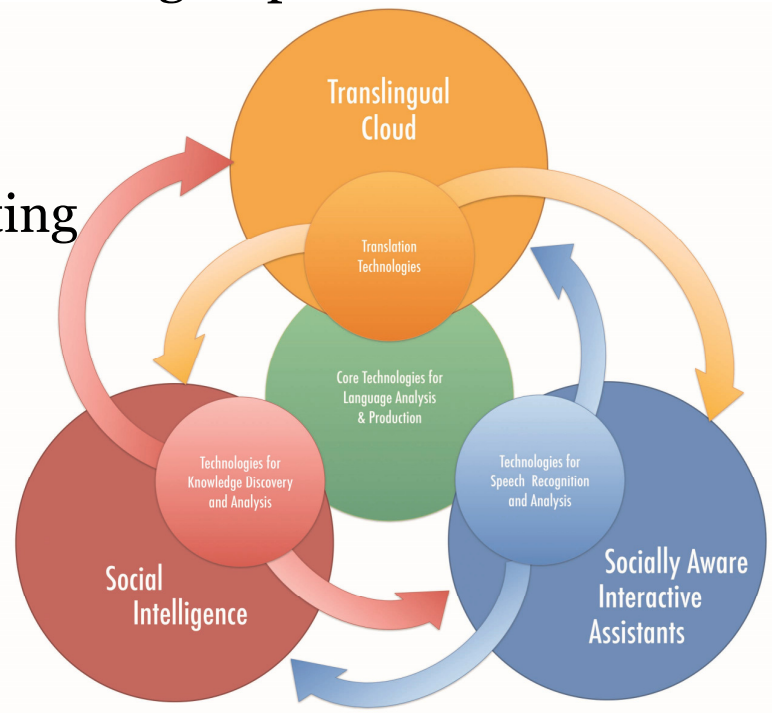
- Better decisions by monitoring social media
- Inclusion of citizens into collective decision processes
- Opinion formation, consensus building, decision making
- Evolution of new solutions
- New forms of democracy: e-democracy, massive participation, transparency
- Dialogues and debates across language boundaries and across parties, political alliances, social classes
- Documented transparent decision processes

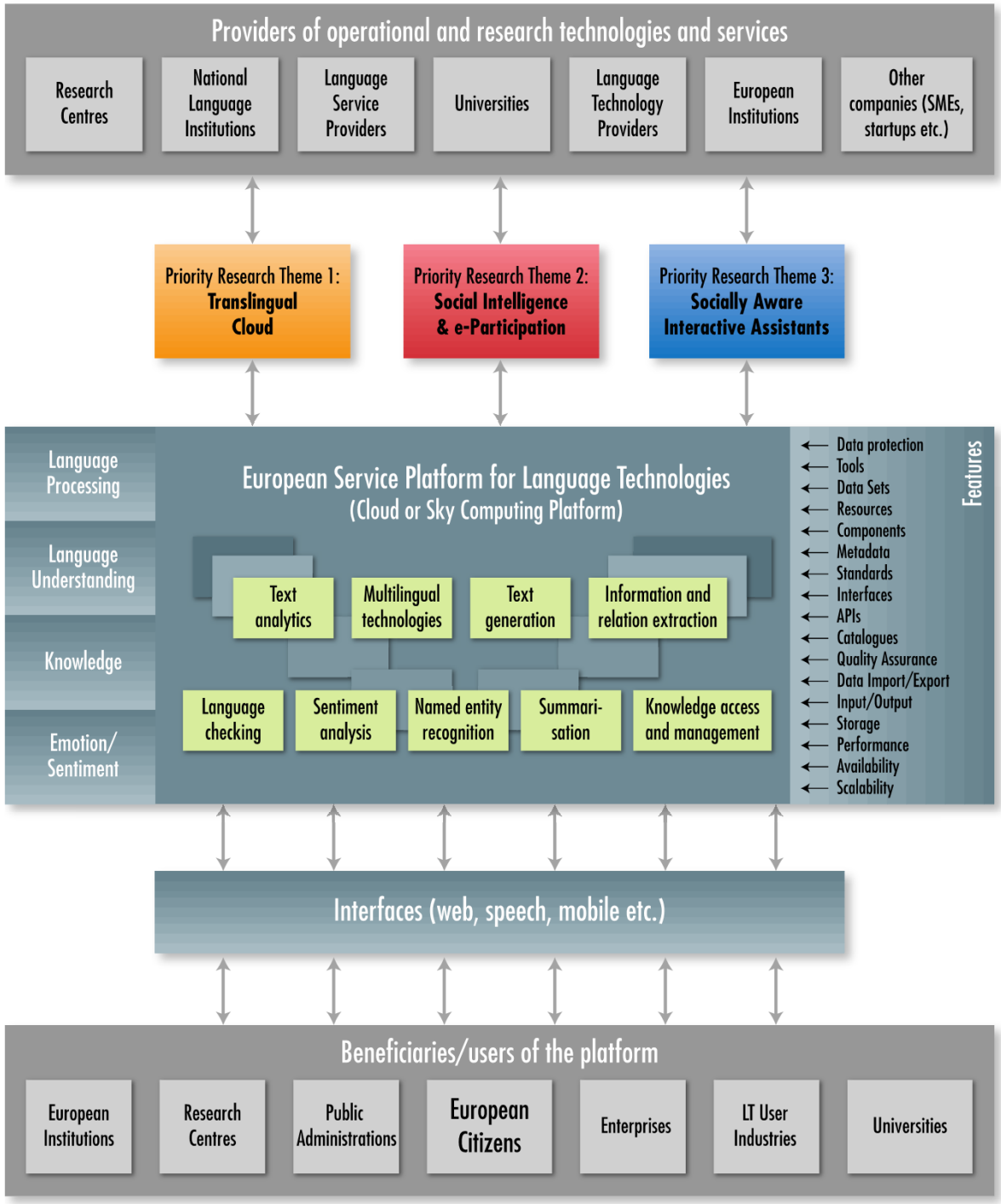



Priority Research Theme 2: Social Intelligence and e-Participation
Mapping large, heterogeneous, unstructured volumes of online content to structured, actionable representations
From shallow to deep, from coarse-grained to detailed processing techniques
Making language technologies interoperable with knowledge representation and the semantic web
"Semantification" of the web: tight integration with the Semantic Web and Linked Open Data
Services and Technologies:
Intelligent analysis of web content, especially social media, comments, blogs, forums
Detection and cross-lingual analysis of decision-relevant information
Multilingual, problem-specific decision support
Text analytics (named entity recognition, event recognition, relation extraction, sentiment analysis and opinion mining including the temporal dimension)
Syntactic, semantic, rhetorical analysis and text structure identification
Resolution of coreference or modality cues
Extraction of semantic representations from arbitrary online content
Clustering, categorising, summarising, visualising discussions and opinion statements
Applications:
Technologies for decision support, collective deliberation and e-participation
Public discussion platform for Europe-wide deliberation on pressing issues
Visualisation of social intelligence data and processes; modeling evolution of opinions
High performance web-scale content analysis technologies
Events/trend detection and prediction
Make use of the wisdom of the crowds
Improved efficiency and quality of decision processes
Unleashing social intelligence by detecting and monitoring opinions, demands, needs and problems
Target groups: European citizen, European institutions, discussion participants, companies
Understanding influence diffusion across social media

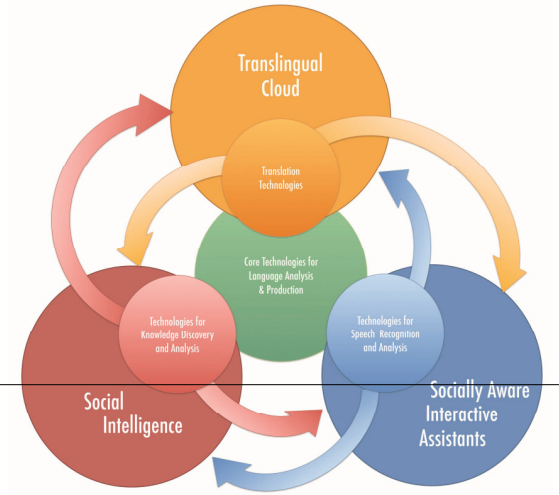
PT3: Interactive Assistants

- Interact naturally with humans
- Use language in connection with other modalities
- Interacting naturally both with and inside of groups
- Multilingual
- Self learning
- Adapts to environments – social/meeting




Providers of operational and research technologies and services
Research Centres
National Language Institutions
Language Service Providers
Universities
Language Technology Providers
European Institutions
Other companies (SMEs, startups etc.)
Priority Research Theme 1: Translingual Cloud
Priority Research Theme 2: Social Intelligence & e-Participation
Priority Research Theme 3: Socially Aware Interactive Assistants
Language Processing
Language Understanding
Knowledge
Emotion/ Sentiment
European Service Platform for Language Technologies
(Cloud or Sky Computing Platform)
Text analytics
Multilingual technologies
Text generation
Information and relation extraction
Language checking
Sentiment analysis
Named entity recognition
Summari-sation
Knowledge access and management
Features
Data protection
Tools
Data Sets
Resources
Components
Metadata
Standards
Interfaces
APIs
Catalogues
Quality Assurance
Data Import/Export
Input/Output
Storage
Performance
Availability
Scalability
Interfaces (web, speech, mobile etc.)
Beneficiaries/users of the platform
European Institutions
Research Centres
Public Administrations
European Citizens
Enterprises
LT User Industries
Universities

Core Resources & Technologies



META-NET

Conclusions and Next Steps

Conclusions and Next Steps

- Europe is extremely interested in and passionate about its languages.
- Our Strategic Research Agenda for LT research and innovation can put Europe ahead of its competitors in this technology area.
- Provides useful and attractive solutions to European society, at the same time creating huge business opportunities for European industry.
- Now is the time to move forward with a continent-wide, systematic push and to invest in strategic research. A modest investment is required.
- We are very confident that we can help build applications that break down language barriers in Europe and beyond.
- This push will generate a *countless* number of opportunities.

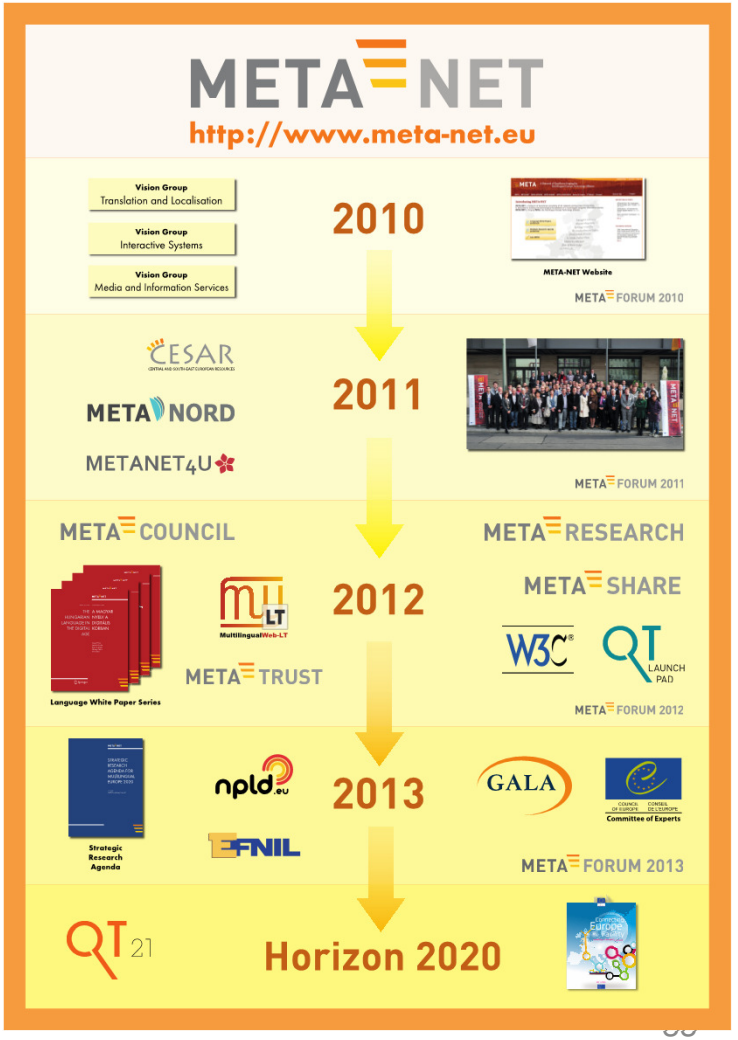
Thank You!

META NET

Acknowledgements: This work would not have been possible without the dedication and commitment of our colleagues Georg Rehm, Aljoscha Burchardt, Kathrin Eichler, Tina Klüwer, Arle Lommel, Felix Sasaki and Hans Uszkoreit (all DFKI), the 60 member organisations of the META-NET network of excellence, the ca. 70 members of the Vision Groups, the ca. 30 members of the META Technology Council, the more than 200 authors of and contributors to the META-NET Language White Paper Series and the ca. 200 representatives from industry and research who contributed to the META-NET Strategic Research Agenda.

jjudge@computing.dcu.ie

ie.linkedin.com/in/judgejohn/

http://www.cngl.ie

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