Strategic Research Agenda for Multilingual Europe 2020

Priority Theme 1: The Translingual Cloud

Jan Hajič

Institute of Formal and Applied Linguistics
Charles University in Prague
Czech Republic

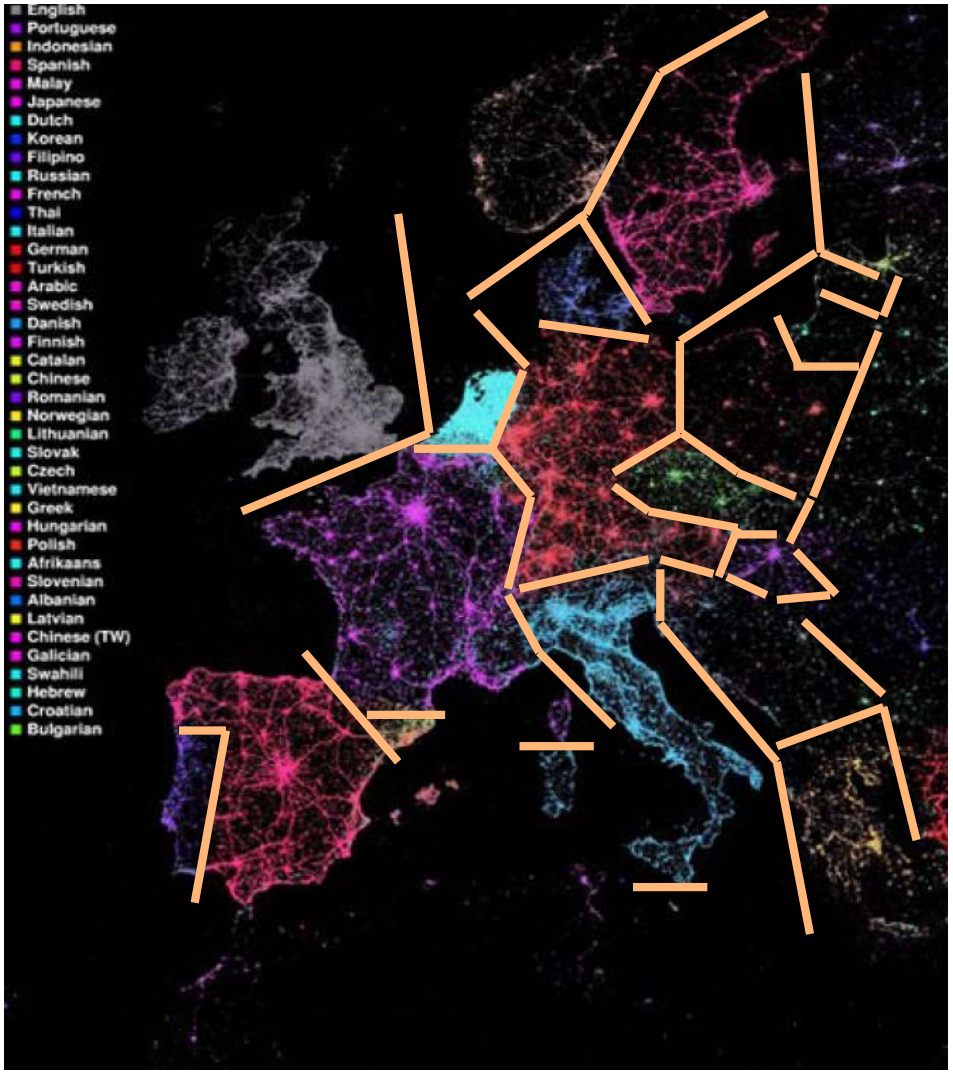
January 25, 2013 – Berlin, Germany

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 of the Priority Theme 1 Translingual Cloud

- Solutions for the EU Society and the Citizen
  - No language barriers – written or spoken, ubiquitous presence
- Novel Research Approaches
  - Goal: High Quality MT - all languages, all situations, all domains
- Solutions and Technological Implementation
  - Translation in the cloud / as a service
- Impact
  - Seamless multi- and cross-lingual applications → business opportunities
- Organization of Research
  - Infrastructure + (hybrid) funding + interdisciplinarity + evaluation

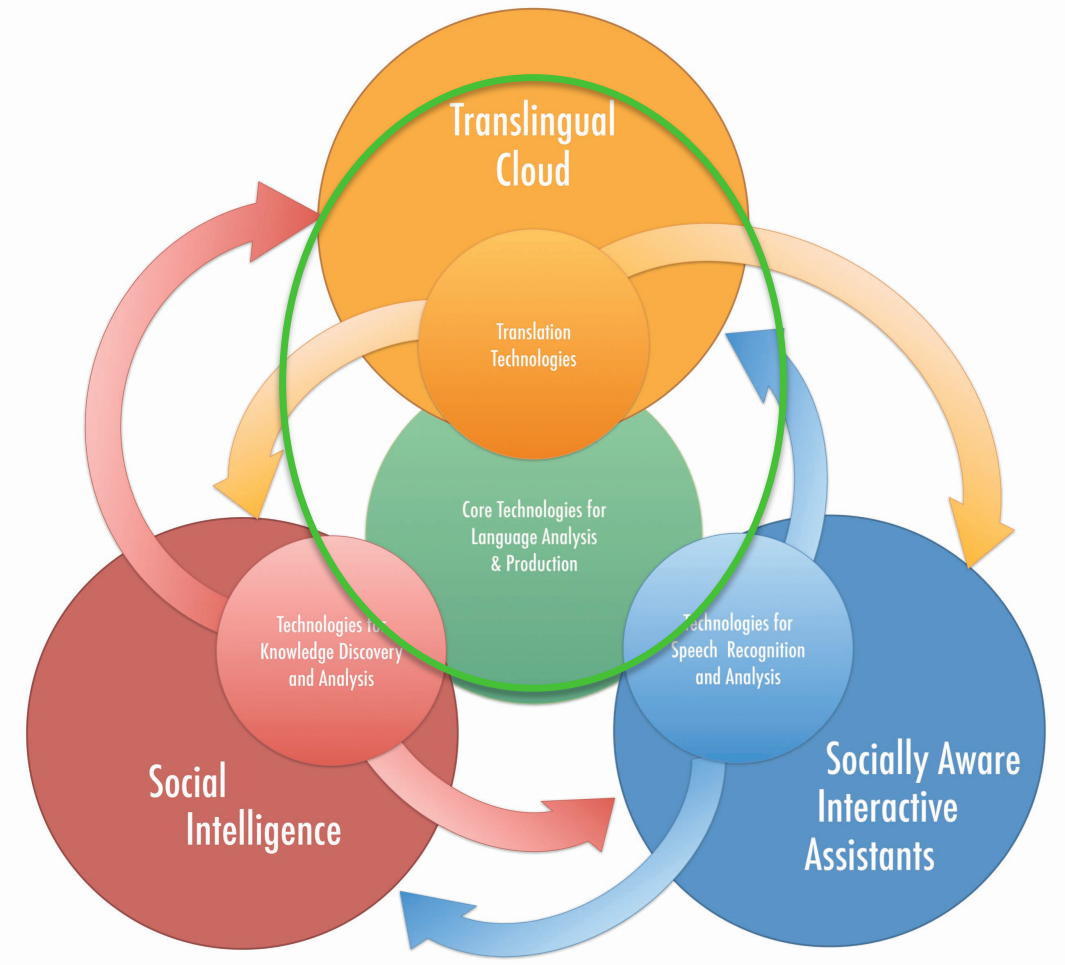
The Situation: Society



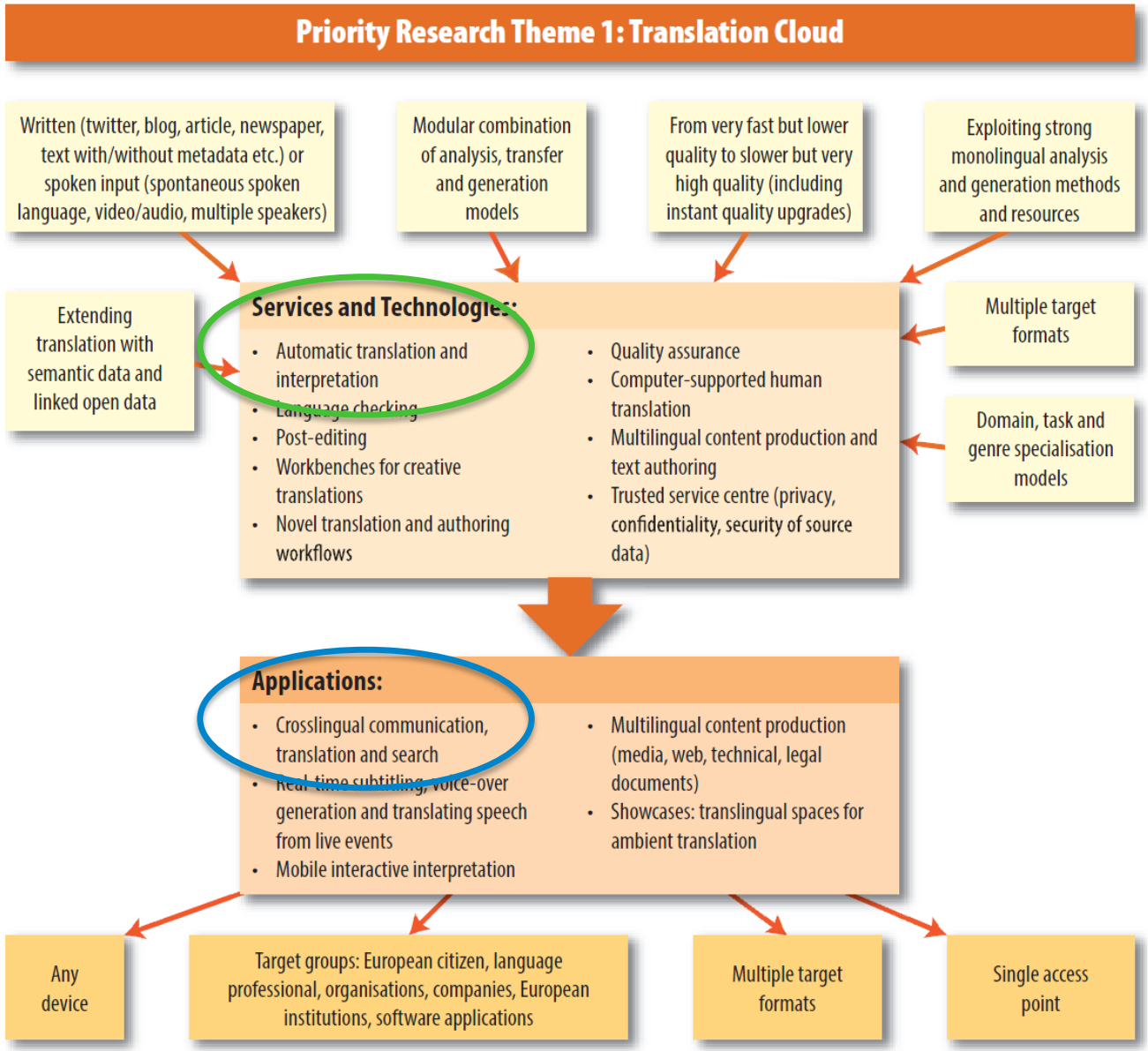
(← Twitter language map, EU)

- Language communities mimic borders (almost)
- National languages prevalent (no surprise)
- (Language) borders limit:
  - Discussion
  - Commerce
  - Flow of ideas
  - Relations, friendships
  - Mobility
  - Society development
  - Research
- Danger of
  - Digital extinction of most languages

The Translingual Cloud in Context



Translingual Cloud – The Scheme



The Situation: Technology

- 30 European languages compared (Language White Papers)

	Support: excellent	good	moderate	fragmentary	weak/none
Machine Translation		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
Language Resources		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

Translingual Cloud Roadmap

- Current State - each topic planned in three phases
  - Immediate availability of translation
    - 2013-2014: data, tools, novel evaluation, identify gaps
    - 2015-2017: towards HQMT systems & showcases, adaptation, non-EU lang.
    - 2018-2020: first applications of HQMT deployed, novel workflows
  - Multilingual multimedia content creation and delivery
    - 2013-2014: combined audio/text/video/image analysis, evaluation metrics
    - 2015-2017: prototype applications (public, A/V industry), new languages
    - 2018-2020: large-scale applications, online services, non-EU lang.
  - Cross-lingual content management, linked open data
    - 2013-2014: publication of resources, multilingual ontologies/data
    - 2015-2017: tools and services for combining free text & data for HQMT
    - 2018-2020: cross-lingual seamless access to linked open data
  - Content analytics
  - (A)synchronous interpretation (speech translation)
  - Translingual collaborative spaces

Beyond Current Roadmap

- Application areas
  - Getting input from QTLaunchPad project
    - Corporate use ("export", localization)
      - Large companies (automotive, IT), SMEs (different approaches, tools, services)
      - Outbound documents, internal use, customer-generated (inbound)
    - Public use (governments at all levels, public services)
      - Legal, employment, social services, eLearning, personal communication, cultural heritage
    - Health Care / Medical (improving care, lowering cost)
      - Emergencies, delivery of information to public / professionals, health care system interoperability, health care provider cooperation
    - Media (access, creation, delivery)
      - TV, film, internet; online services (subtitling, voiceover), news delivery, eLearning
- Foundational Technology
  - Necessary progress – basis for all applications
    - New projects
- Data Resources
  - Filling identified gaps, novel areas (multimedia – not even in WP tables)

Roadmap for Translingual Cloud (1)

- Foundational technology, evaluation, targeted prototypes

Research Priority	2013-2014	2015-2017	2018-2020
MT as a service – immediately available, cheap, appropriate quality	Monolingual ("deep", semantic-aware) tools, inclusion in MT, methodology, machine learning for MT, context	Complete HQMT systems (hybrid), evaluated by novel metrics; deeper work on EU, start non-EU languages	Deployment of HQMT systems, services in the "Cloud"; targeted and tailored APIs; adaptation tools; more non-EU languages
Evaluation methodology supporting High Quality MT	Devise novel metrics, correlated with subjective meaning preservation, fluency	Incorporate feedback from research systems, develop datasets supporting new metrics, best practices	Provide evaluation infrastructure, structured to areas, applications, languages
Targeted Prototypes and Applications for the Corporate and Public use	Targeted (narrow) domains, show feasibility (medical, car industry, public services, ...)	First applications with full MT workflow integrated (esp. in public services), collecting feedback	Full prototypes in many areas, using the Cloud services, added languages (export-oriented)

Roadmap for Translingual Cloud (2)

- Multimedia and interpretation, ambient translation

Research Priority	2013-2014	2015-2017	2018-2020
Delivering multimedia content in any language	MT for combined text, audio, video, image content, in languages with enough resources, novel Machine Learning for MM	Prototype applications in MM domains (governmental, entertainment), A/V archive search and delivery, MT for user-generated content	Deployment of large-scale applications for MM content, online subtitling/CC services, new languages of EU business interest
Synchronous and asynchronous interpretation, collaborative spaces	Improving ASR in low-performance areas, speech translation for lectures, presentations, languages with existing resources, metrics for evaluation of interpretation	Novel areas of application in selected domains (synchronous interpretation systems), context-aware methods or ASR and spoken language translation, open-space translation and interpretation	Deployment of first systems as a service for interpretation and spoken material translation (both synchronous and asynchronous), dedicated spaces with translingual services

Roadmap for Translingual Cloud (3)

- Crosslingual knowledge management, LOD; resources

Research Priority	2013-2014	2015-2017	2018-2020
Multi- and Cross-lingual knowledge management and structured data, content analytics	Linking multilanguage text and audio resources with data and metadata (ontologies, structured data, wikidata, ...), information extraction from multilingual sources, using languages with enough resources. New evaluation methods.	Cross-fertilization between KR and MT research (novel techniques combining text and structured data for MT); extracting knowledge from text and speech. Create test data and use cases (consistency, cultural aspects, timeline, ...)	Services and applications for accessing and using ontologies, other structured data in many languages; services and systems for updating structured resources (ontologies, databases)
Resources	In coordinated fashion (META-SHARE, other repositories) identify gaps (languages, technologies, volume)	Acquire, annotate, distribute LR data for MT in all areas of interest in the Roadmap	Continue collection of identified gaps, non-EU languages of interest, new areas, new MT technology needs

Q/A

Thank you very much!

office@meta-net.eu

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