Empowering a Multilingual Continent

Technologies and Language-Centric AI for Language Equality in Europe

This is a Call for Action, addressed to the European Institutions, issued by the European Language Technology community, as represented by LT-Innovate, the Association of the Language Technology Industry (200+ member companies), and the Multilingual Europe Technology Alliance, META-NET (a network of 60 research centres in 34 European countries).

Languages and the Cultural Integrity of Europe

- Our citizens use language to express, store, access, share, manipulate, interpret and search massive amounts of information. They do this in written text, spoken language, audio-visual materials and dialogue in almost all of their social, economic, business, political, legal, scientific and cultural interactions. Digital technologies facilitate many of these interactions.
- Multilingualism is at the heart of the European idea. The EU has 24 official languages and many additional ones (regional and minority languages/dialects, languages of immigrants and trade partners).
- The 24 official EU languages are granted equal status by the EU Treaties and the Charter of Fundamental Rights in the EU.
- However, language barriers still hamper cross-lingual communication, human mobility and the free flow of knowledge, ideas, commerce, administrative, cultural and political exchanges in Europe.
- We are also facing a widening technology gap between widely-used and lesser-used languages of Europe that deepens the digital divide for economically less powerful linguistic communities.

Artificial Intelligence has Language at its Core

- Over the coming 10-15 years, artificial Intelligence (AI) will transform every domain of industry and society, from construction & manufacturing to commerce & retail, education & healthcare to legal & finance, defence & security to entertainment & tourism, etc.
- Language Technology is one of the most important drivers behind the current boom in AI.

- Language Technology is also at the very heart of the cognitive revolution redefining the interaction between humans and machines; humans will speak with AI-empowered machines and robots.
- Deep Natural Language Understanding is the ultimate goal of AI.
- Currently, almost all developments are in the English language only. Most European citizens and their countries will be excluded from these game-changing technologies if we do not ensure that they have multilingual capability.

A World-class Research Landscape

- Europe has a world-class research landscape in Language Technology, Computational Linguistics and related fields.
- Through continued investments in FP6, FP7 and H2020, the EU has contributed to this situation.
- AI and LT research keep on making important technology breakthroughs.
- In recent years, there have been massive investments in AI R&D and applications outside Europe (mostly in the USA and Asia).
- In Europe, the current funding schemes are unbalanced, erratic and too small. There is a real danger for Europe to lose touch with leading edge research and novel, potentially paradigm-shifting global developments unless the investment in Europe is not only sustained but increased through a combination of public funding mechanisms and private corporate investment & venture schemes.

A Thriving Industry Hampered in its Expansion

- Europe has a large number of innovative established SME, as well as start-up and scale-up companies working in this area. For many of them, an optimal outcome of their venture is to be acquired by global players (who are mostly of non-European origin).
- A few mid-tier European companies are successfully scaling up and opening commercial & research branches across Europe and overseas and are acquiring and investing in selected start-ups. An ecosystem that builds on the specificities of the European LT industry should be developed for European LT to grow.
- Persistent ‘fragmentation’ is largely due to the absence of a truly European Single Market on the one hand and the absence of a performing European LT Infrastructure that bridges the national language silos on the other hand.
- The European LT ecosystem is B2B-centric: many of the clients that purchase language technologies or services come from specific vertical markets such as software, retail, manufacturing, automotive, healthcare, administration, culture, etc. This vertical focus should be strengthened and transformed into a competitive advantage.

Technologies for Multilingual Europe – Challenges

- While the 24 EU languages enjoy equal status, digital extinction of at least 21 of these languages is a real danger.
- As long as the Digital Single Market is not fully multilingual, Europe will be composed of 20+ isolated markets!

- Citizens should be empowered to use their mother tongues and/or the language(s) of their choice with any digital technology.
- Trust in the media needs to be restored (e.g. with regard to online propaganda and misinformation) by enabling access to quality news from diverse cultural and linguistic sources.
- With the ubiquity of ICT and multilingual nature of Europe, individual end users and all types of companies have a big need for high-quality, high-coverage, precise, robust Language Technologies.
- Due to the persistent market fragmentation, the European LT ecosystem is unable to satisfy the B2C demand across the continent which, in consequence, often tends to be monopolised by large global players.
- The European language challenge cannot be abandoned or outsourced. Europe must not make its digital infrastructure dependent on non-European solutions (cf. GALILEO vs. GPS and GLONASS).
- The European language challenge should become a priority not only for the European Institutions, but for Members States and regions, e.g. by being declared an obligatory key "smart specialisation" topic to effectively foster synergies between European, national, bi- or multi-national and regional funding instruments.
- Finally, it might be beneficial to explore innovative ways of reversing the usual process from Research to Development to Product by early deployment of technology that is “good enough” (rather than perfect) and then involving user interaction (crowds, citizens) in improving language technologies. Unlike other ‘resources’, language is an asset of the people.

A Myriad of Opportunities around Technologies and Multilingualism

- Europe has the unique opportunity, through the deployment of its technologies, of leveraging its large market into a fully multilingual (or, rather, language-neutral) space, turning a collection of fragmented markets into a truly unified and inclusive Europe, while supporting our rich and diverse linguistic heritage.
- An Opportunity for European Society: Enable all European citizens to communicate with each other and to use digital technologies and connected devices in their own languages.
- An Opportunity for the Media: Address the increasing social, political and commercial relevance of online content and communication (online propaganda, misinformation campaigns, fake news etc.).
- An Opportunity for the Public Sector: address Europeans in the language they speak, enable efficient Pan-European and cross-border services.
- An Opportunity for the Digital Single Market: Enable European SMEs to grow and scale up by realising the *Multilingual* DSM.
- An Opportunity for our Languages: Future-proof our languages by addressing the threat of digital extinction.
- An Opportunity for the Internet of Things: High-quality spoken language interfaces for billions of connected devices – from wearables to household appliances to cars to whole manufacturing lines – in all European languages.
- An excellent set of opportunities for Europe, European research, European education, European industry, European innovation, European workforce and European culture!

- Address the ever-increasing brain drain of young researchers and high potentials leaving Europe by creating attractive incentives for them to pursue their research and build their careers here.
- Language Technology *for* Europe *made in* Europe is the key for European cross-border and cross-language communication, economic growth and social stability.
- We can move Europe into the pole position in this field and deploy the model in other multilingual societies.
- Europe needs European LTs to avoid reliance on non-EU market monopolies. Europe has to develop its own LT infrastructure, based on a close partnership between research, innovation and industry.

Important Research & Innovation Opportunities:

- All recent policy papers (see Further Reading below), including the most recent study "Language Equality in the Digital Age – Towards a Human Language Project" commissioned by the European Parliament (March 2017), emphasise the necessity of a large-scale language programme, tailored to Europe's needs and demands. Here are some key R&I directions that they have identified:
    - **Language, Machine Learning and AI:** Advances in AI together with increased availability in computing power and data have massively improved LTs and facilitated the integration of modalities and sensor input (text, speech, vision). LT needs close collaboration with ML/AI to tackle challenges including explainability and stability (repeatability over various runs), learning from small data sets, finding good model architectures, fast correction of models etc.
    - **Language and Knowledge:** Many large collections of (mostly factual) knowledge extracted and curated manually and semi-automatically are available online. The challenge is to find optimal ways of making important structured data/knowledge available to ML-based LTs.
    - **Interoperability:** While there has been a trend towards end-to-end trained systems, modularity of systems and resources will remain important in many large applications and may positively impact on, e.g., explainability of complex systems. The challenge here is to develop technology support for interoperability.
    - **Language Data of Industrial Relevance:** Raw and labelled data (from and for industry, public and research) needs to be collected, assessed, processed and made easily available, for continuous integration into our technologies. Regarding IPRs, a broad, unified TDM exception which includes industrial exploitation is necessary to compete with companies working under the "fair use" principle.

A Strategic Investment that Europe Cannot Afford not to Make

- Invest in four pillars: To realise the vision of a multilingual continent fully empowered by language technologies, we suggest expanding two existing and initiating two new strategic investments.
- **Pillar 1: Basic Research – Towards Deep Natural Language Understanding**
    - We recommend to establish the "Human Language Project" (HLP) as one of the next EU FET Flagships in 2021.

- The European language technology and language-centric research landscape is already world-class but needs to be united under a shared umbrella to arrive at the next scientific breakthrough.
- The project proposal "Human Language Project Preparation" (HLP Prep) will be submitted in Sept. 2018. If successful, this 12-months project will develop a full EU FET Flagship proposal, starting in early 2019.
- The scientific goal is to reach Deep Natural Language Understanding by the year 2030. This is the ultimate goal of LT: systems that can learn, reason, interact and explain themselves, and that can do so robustly and reliably across languages, modalities, platforms and cultures.
- If Europe stops investing into basic and applied research, it risks losing touch with research. Instead, we suggest to invest substantially in this area so that LT and Language-Centric AI finally emerge as a truly European set of technologies and applications that will turn out to be a truly European USP around the globe.
- One cannot periodically switch on and off research and innovation in this important technology sector. Underfunding risks the EU falling behind the global competition. European LT research is competitive at the highest level, if research strengths across Europe are bundled.
- *Suggested instrument:* EU FET Flagship

- **Pillar 2: Multilingual Digital Public Services**
  - The Connecting Europe Facility SMART programme that initially supported the creation of an EC machine translation system to make selected Digital Service Infrastructures multilingual needs to be continued and expanded.
  - By now CEF SMART also attempts to attract business interest through procurement actions. These actions should be simplified so as not to be dissuasive for SMEs.
  - Public procurement of the European multilingual infrastructure should serve as a major driver for the growth and consolidation of the European LT industry.
  - A strong pan-European community has been built and includes national administrations as valuable public data providers. Hundreds of valuable data sets (PSI) have been collected already. We should continue to make these available to research and industry and their use in citizen-centric services should be encouraged.
  - *Suggested instrument:* CEF SMART
- **Pillar 3: European LT Business Space**
  - A first seed Innovation Action project is expected to be launched at the end of 2018, supported through the Horizon 2020 ICT-29 call ("Multilingual Next Generation Internet – European Language Grid").
  - This first step at organising and fostering the collaboration between the various components of the LT ecosystem needs to be scaled-up substantially and maintained throughout the duration of the Horizon Europe programme to reap benefits through a coherent and sustained effort. Europe needs a basic European Language Infrastructure for natural language processing. All language processing applications (search, mining, writing, speech, translation, etc.) depend on such an infrastructure. The infrastructure should provide the basic functionalities required to process unstructured content. Through APIs it should provide basic LT services

such as tokenization, stemming, part of speech tagging, named entity detection, identification of measurements, currencies, formulas, etc. for all languages, in the same basic quality, under the same favourable terms. This is part of the suggested European Language Grid, integrates results of previous CEF, H2020 and FPs, but it might be extended by other projects and sources.
    - It should be progressively transformed into an Innovation Hub in which technology providers, potential customers and applied research collaborate on joint projects. This effort should be combined with a mechanism to ascertain industry needs on an ongoing basis.
    - *Suggested instrument:* Corresponding calls for Innovations Actions and Coordination and Support Actions under the umbrella of Horizon Europe, possibly also under (the more narrowly focused) CEF SMART.
- **Pillar 4: Innovation, Integration and Deployment of AI-Centric Language Technologies**
    - Innovation projects should be clustered around domain specific vertical platforms (e.g. media, automotive, finance, pharma, healthcare, tourism, etc.)
    - Flagship technology innovation, integration and deployment projects should be launched to introduce newly developed technologies to the market as soon as possible. Large industrial technology users should be incited to co-invest into these flagship projects in return for solutions to specific problems in their sectors.
    - Framework agreements should be worked out per domain specific platform governing all relevant aspects of collaboration (work sharing, liability, IP ownership and licensing, co-financing, etc.).
    - *Suggested instrument (innovation and integration):* Corresponding calls for Innovations Actions under the umbrella of Horizon Europe.
    - *Suggested instrument (integration and deployment):* Corresponding funding should be made available through the AI-component of the emerging Digital Europe Programme.
    - *Suggested instrument:* Innovation Actions under the umbrella of Horizon Europe.

Brussels, Prague

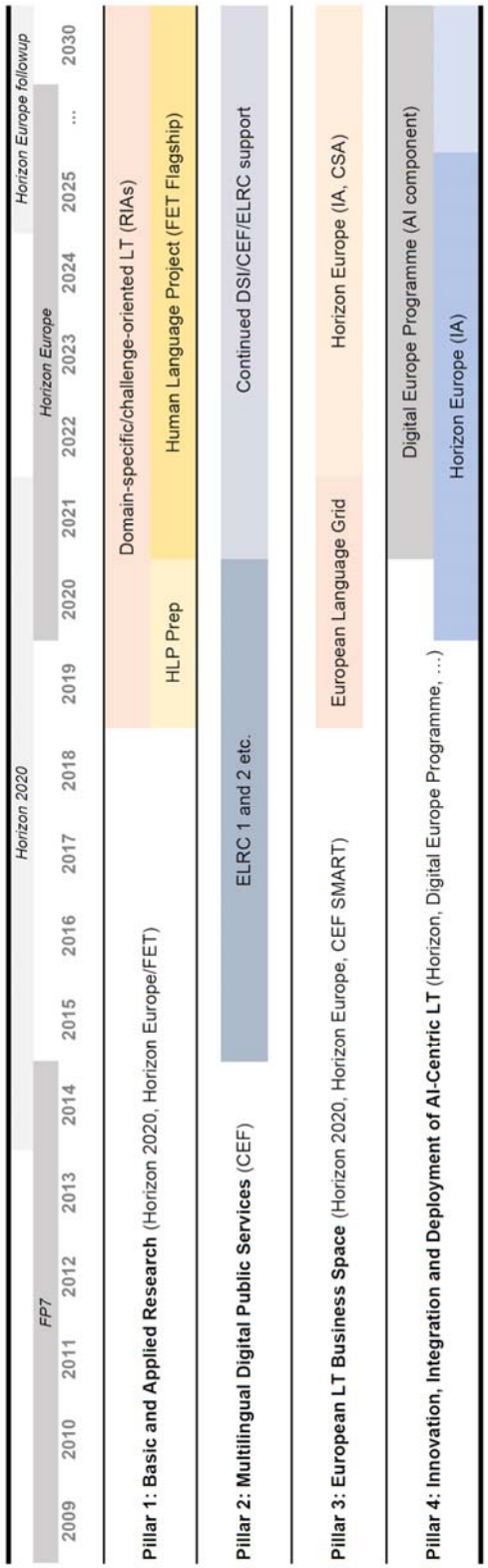
July 13, 2018

Jochen Hummel
Chairman of the Board of Directors, LT-Innovate

Jan Hajič
META-NET Chair of Executive Board

Appendices

Timeline



Further Reading

- European Parliament Draft Report on *Language Equality in the Digital Age* (2018)
- Cracking the Language Barrier Strategic Research and Innovation Agenda *Language Technologies for Multilingual Europe: Towards a Human Language Project* (2017)
- European Parliament Science and Technology Options Assessment (STOA) *Language equality in the digital age: Towards a Human Language Project* (2017)
- LT-Innovate *Assessment of the State of the EU Language Technology Sector and EU Policy Recommendations* (2016)
- Riga Summit statement *Multilingual Europe: the Crowning Touch to the Digital Single Market (2016)*
- META-NET *Strategic Research Agenda for Multilingual Europe 2020* Springer (2013).
- Rehm & Hegele *Language Technology for Multilingual Europe: An Analysis of a Large-Scale Survey regarding Challenges, Demands, Gaps and Needs* (2018)
- Wacker & Joscelyne *Improving Collaboration Between the European Language Technology Industry and Research: A New Framework for Supply and Demand* (2017)
- Vasiljevs et al. *European Platform for the Multilingual Digital Single Market: Conceptual Proposal* (2016)
- Maegaard et al. *Providing a Catalogue of Language Resources for Commercial Users* (2016)
- Rehm et al. *Fostering the Next Generation of European Language Technology: Recent Developments – Emerging Initiatives – Challenges and Opportunities* (2016)

Contributing Authors from META-NET

Josef van Genabith, Jan Hajič, Georg Rehm, Andrejs Vasiljevs