LT-INNOVATE.EU

THE FORUM FOR EUROPE'S LANGUAGE TECHNOLOGY INDUSTRY

December 2016

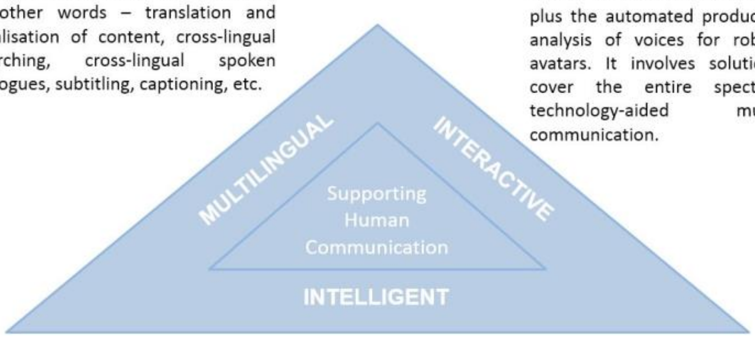
ASSESSMENT OF THE STATE OF THE EU LANGUAGE TECHNOLOGY SECTOR AND EU POLICY RECOMMENDATIONS

1. STATUS OF THE HLT SECTOR

- What is your perception of the current status of the HLT sector in the EU?
  - **HLT is to be understood as a combination of three sets of technologies (in short, "translation", "speech" and "analytics"):**

Multi- and cross-lingual processing covers all of the contexts in which the presence of multiple languages in the content communication chain requires language-specific solutions. In other words – translation and localisation of content, cross-lingual searching, cross-lingual spoken dialogues, subtitling, captioning, etc.

Interactive communication covers all processing and analytic operations of spoken language (syntax, semantics and pragmatics including emotion, tone, and similar features) in face to face or remote interaction contexts, plus the automated production and analysis of voices for robots and avatars. It involves solutions that cover the entire spectrum of technology-aided multimodal communication.



Intelligent content covers all processing and analytic operations that use natural language processing (syntax & semantics) to parse, understand, link, categorise and leverage text content found in any media (visual or textual documents), often bundled inside other applications. It involves solutions that automate and accelerate the production, communication, reception and comprehension of all these content-centric processes.

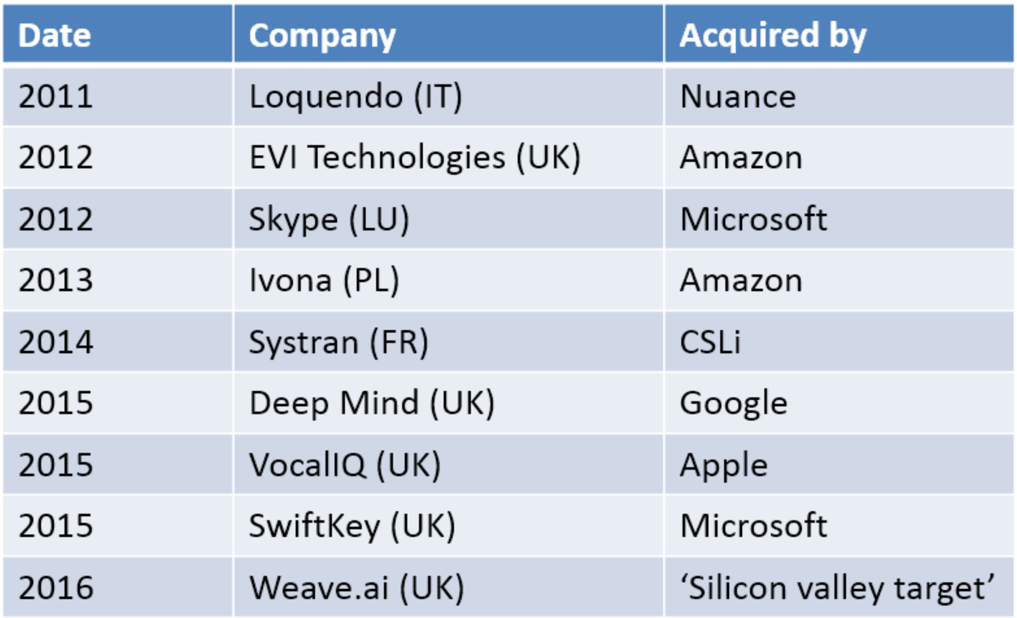
  - **HLT is an opportunity sector.**

Core LT Markets	in € by 2015	in € by 2020
Intelligent content	6.5bn	20bn
	Intelligent big data	
Interactive communication	8bn	15bn
	Human machine communication	
Multi-and cross-lingual processing	12bn	30bn
	Massification of translation	
TOTAL	26.5bn	65bn

  - **Europe has many companies and university groups working in this area. However, there are only a handful of successful HLT companies in Europe, and a huge long tail of start-ups and partly not very viable activities masquerading as HLT companies – most of them attached to Language Service Providers (LSPs).**
  - **Europe is a world leader in translation technology and has strong positions in speech and analytics, but its leading providers may not have the resilience to scale-up to become global actors.**
  - **HLT are key enablers for mega trends such as eCommerce, Big Data, Artificial Intelligence (AI), and eGovernment interoperability. However, most HLT products have to be developed individually for every single language. This creates fragmentation in these key technologies and makes it hard for scale-up companies to be successful.**
  - **Language is the main barrier for the Digital Single Market. Game-changing IT products work only in English. Most European countries, their industries, and their citizens are basically excluded from the data revolution. Without an interoperable eGovernment, the EU will continue to attempt solving crises top-down and fail.**
  - **The biggest translation organization in the world, the European Commission, has no budget and process for deploying market-ready efficiency solutions. In-house developments destroy the market for superior and cheaper industry solutions.**
  - **Excessive EU policy focus on minority languages obfuscates the fact that, if solutions exist for main languages (i.e. those that have the biggest purchasing power), minor languages will benefit from them.**

- What is the situation compared to other countries/continents?
  - **Natural Language Processing (NLP) enjoys a huge renaissance in the US due to the breakthrough in Artificial Intelligence (AI).**
  - **Unlike the US, Europe has very few large billion $ software companies to invest (and promote an ecosystem) around HLT.**
  - **The promotion of university or public institution funding for open-source software by the EU is good for the services industry but effectively killing the HLT start-up or small company ecosystem in Europe.**
  - **A few successful EU HLT companies in Italy (Expert System), Spain (Inbenta, Unbabel), the UK (Creative Virtual) and Germany (Acrolinx) have opened up major commercial & research branches in Silicon Valley. Instead of being acquired, they have rebranded themselves as global companies.**
  - **A few innovation-oriented European mid-tier firms are acquiring and investing in selected HLT SMEs, like Bertin Technologies, in France (acquired Vecsys in 2011, AMI software in 2016), Almaviva in Italy (acquired Pervoice in 2014). Instead of letting European technologies get acquired by US firms, they push further European technological advances and are credible alternatives to main US providers when bidding on large projects in Europe and outside Europe.**

- **Large foreign (mostly US-based) companies have been systematically acquiring promising EU companies, mainly UK-based, in the HLT field. Very few of these companies have been acquired by large EU firms. Meanwhile, big US digital companies have been 'milking' EU innovation funds.**



Date	Company	Acquired by
2011	Loquendo (IT)	Nuance
2012	EVI Technologies (UK)	Amazon
2012	Skype (LU)	Microsoft
2013	Ivona (PL)	Amazon
2014	Systran (FR)	CSLi
2015	Deep Mind (UK)	Google
2015	VocalIQ (UK)	Apple
2015	SwiftKey (UK)	Microsoft
2016	Weave.ai (UK)	'Silicon valley target'

- **The emergence of strong European alliances (across the 3 HLT segments described above e.g. between speech and semantics/analytics technologies integrating self-learning and deep-leaning) would be desirable.**

- **Unless the situation changes radically in Europe, it will be increasingly difficult for European players to maintain an edge over their competitors.**

- **More and more US companies have understood that the global market is multilingual. Heavy hitters such as Google, Microsoft, IBM, Facebook, Amazon, etc. invest huge amounts in HLT. They are close to becoming dominant players in the field. The public sector (particularly the Department of Defence) procures HLT solutions before they are even fully productized, thus promoting market oriented developments. This is accelerated by the tendency of large European companies and institutions to use "free" US solutions, because they do not understand the value of data.**

- **China has a number of very large digital companies such as Alibaba, Ten Cents and Baidu that have in-house HLT research groups that are closely monitoring foreign HLT solutions for communicating online as well as developing their own solutions. In translation, their language pairs seem to be mostly Chinese and English, and they have made very little effort so far to broaden the scope of their language pairs. But this will change. A lot of Chinese researchers move between the US (and sometimes EU universities) and big Chinese firms or universities, and they all openly share the basic directions of current HLT – statistical translation, "deep" machine learning for speech applications, and artificial intelligence. What makes China important is the size & power of the population of researchers, and size of the native language challenge that they wish to overcome. For example, the Chinese government plans** to introduce codes for some 3,000 Chinese characters as part of a grand project, known as the China Font Bank, to digitize 500,000 characters previously unavailable in electronic form. Until now, only 80,388 characters have been encoded in the international computing standard, Unicode. This is the largest state-funded digitization project ever undertaken. Most specialists consider it completely unnecessary. But it demonstrates massive power and a concern for minority language communities in the country.

    - **Russia is a relatively unknown quantity. It has a strong tradition of mathematical and engineering skills in HLT. The government has tried to strategically promote a "Eurasian" vision of national partnerships that involve many different Asian languages (Kazakh, Khirgiz, etc), but there is little evidence of serious HLT solutions. Yet Russian start-ups and some SMEs are very much part of the multilingual translation technology and data analysis community in Europe and the US (e.g. LogRus, Abbyy).**
    - **India is a multilingual country, in some ways similar to the EU. Very recently under Prime Minister Modi, there has been a drive to digitise local languages, accept the smart phone as a democratic communication device that must adapt to users, and push for massive multilingual government communication & data sharing. India's R&D force is considerable, with strong personnel links to US and UK universities and companies. Expect to see interesting, low-cost multilingual solutions emerging for less-educated Indian communities.**
    - **Qatar among other, less prominent Arab countries, has a well-funded Computing Centre with a strong focus on HLT. It is able to invite lots of European and US guest academics, who are keeping the country up to date on its HLT agenda. Expect innovations in Arabic language technology in the next 5 to 8 years which should contribute to better European language-Arabic translation systems and local data analysis. The market value of Arabic language HLT has never been high (fragmented market, supply chain problems due to security issues, economic hardship, etc.), but there are a lot of security/intelligence and defence requirements**

- How do you expect the sector and the market will evolve in the next years?
    - **The HLT sector and market is likely to grow tremendously in the next years due to the opportunities of intelligent Big Data, human machine communication and "massification" of translation.**
    - **Unless the business climate radically improves in Europe for start-ups and particularly scale-up SMEs, innovation that successfully goes to market is more likely to emerge from the US and Asia than from Europe.**
    - **In the worst case, Europe might end up licencing most of its HLT from foreign (mostly American) IT companies.**
    - **Given the strategic importance of HLT, ignoring the above described challenges (and therefore opportunities) would fundamentally set back the economic development of Europe for generations to come.**

- What are the main challenges for the sector in the EU?
    - **Fragmentation, mainly due to the absence of the Internal Market that the EU has been announcing for 25 years and the largely impermeable language silos that characterise Europe's content markets.**
    - **Lack of Product Management (few software companies focus on the development of products fit for the global markets).**
    - **Lack of investment (HLT does not appear anymore in the current work programme of Horizon 2020 (!), European-scale VC funding is only slowly becoming available…).**
    - **Erratic purchasing behaviour by large public and private institutions (insufficient procurement combined with the "buy IBM = buy safe" effect).**
    - **US competition (particularly strong because of the 4 weaknesses above).**
    - **Lack of awareness at policy making level of the importance of HLT and its contribution to a successful Digital Single Market.**
    - **Flawed European policies around software and R&D (too many "me too" ideas, projects driven by academics, funding too thinly and widely spread, no real European and global impact).**
    - **Lack of Language Resources (LRs) needed for the development of domain-specific automated translation solutions in all European languages (existing LR "repositories" are not usable operationally, particularly not for commercial purposes).**

2. **US PROVIDERS vs EUROPEAN TECHNOLOGY**.

    - Do you think that the EU currently heavily relies on language technologies developed outside Europe?
        - **Yes: Google, Nuance, IBM Watson, Apple, Microsoft, Amazon, Facebook are the heavyweights… dominating the B2C market segments… and increasingly the B2B segments too.**
        - **Some B2B niche and emerging segments are still dominated by European (i.e. national/local) providers e.g.**
            - **Translation support software**
            - **Sentiment analysis or "voice of the customer" solutions (but, as ambitious global customer care projects emerge, corporations/institutions find it difficult to resist the urge to buy from large US reference companies)**
            - **Media monitoring solutions**
            - **Contact centre chatbot solutions**
            - **Defence and security (monitoring) solutions**
    - What are the reasons of this situation?

   - **The failure of European politicians and business leaders to understand that a challenge is an opportunity. We can lament about the huge homogenous US market... or we could invest in delivering a multilingual Digital Single Market and multilingual eGovernment and become the fittest for the non-English global market!**

- What are the medium and long term consequences of this approach?
   - **One of the last opportunities to catch up with American IT will be soon missed.**
   - **Almost all viable software companies will emerge in the US, where the business culture promotes them. Ultimately, some of these will find themselves in monopolistic situations.**
   - **No new blue-chip European companies appear on the horizon (the last one in Germany is SAP, founded in the 70s).**
   - **Investment and asset value is annihilated in Europe in macroeconomic dimensions (Nokia versus Google & Apple!).**
   - **As software is "eating" everything, European industries like automotive, pharma, consumer goods, security & defence, etc., that are still strong for now, will be marginalised and services like media, finance, eGovernment, tourism etc. will underperform.**
   - **At the end of the day, end users and citizens will get suboptimal service.**
- Is it realistic to expect that foreign companies will take care of Europe's unique multilingual setup?
   - **Probably not. They will take care of the main languages only. Citizens living in smaller countries will have to live without the data revolution, artificial intelligence, software agents, decent search, etc. More than ever, there will be the digital haves vs. the have-nots.**

3. **IMPACT**.

- Where do you think that HLT could have a more profound impact in the EU?
   - **The EU is in an existential crisis since the huge current challenges can only be solved by working closer together, but citizens are afraid of losing their way of working and living, losing their identity. Nobody proposes an answer to this conflict (least of all the current EU leaders). The British voters have opted out.**
   - **The answer is cross-language interoperability. We can make systems interoperable while maintaining their local differences, but we can only make content interoperable through HLT. Without the latter there will be no cross-border data flow.**

Hence, the EU's priorities should be - in order of decreasing importance:

   - Creating a truly integrated Digital Single Market (e-commerce, business trades) that will be particularly relevant for SMEs.
      - **The (Digital) Single Market is a sine qua non to enable European companies to compete globally.**

        - **There will be no Digital Single Market without addressing language.**
    - Providing effective multi-lingual public e-services while facilitating the access to public information (documents, regulation, open data) to all European citizens regardless of their languages at the European, national, regional and local levels
        - **There is only one EU institution which has achieved this for all EU languages, the European Union Intellectual Property Office (EUIPO). Its success has saved Europe billions. Its investment and consequent use of HLT has made it global leader for Intellectual Property.**
    - Fostering the European construction and reinforcing a common European citizenship identity in the multilingual Europe – Increasing cultural exchanges in Europe.
        - **Namely via natural access to more diverse multi-country contents and documents, which will be indexed and analysed together, whatever their source language.**
    - Increasing transnational mobility and improving labour market - Improving European migration policies
        - **If the above happens, this will follow quasi automatically.**
    - Others

4. **CURRENT POLICIES**.

    - Which language and language-related policies (regional level, Member State level, and European level) need to be promoted, changed or adapted in order to foster and support equal LT development for all languages?
        - **The European Commission (EC) should supplement its Digital Single Market strategy with a policy paper on "overcoming the language barriers" taking into account ideas submitted by HLT stakeholders since 2012 (briefly summarised below under 'Deployment Policies').**
        - **The automated translation core platform of the CEF Programme (so called 'CEF.AT') should be expanded (both in scope and in funding) so as to be able to develop and maintain a European Language Infrastructure as described below.**
        - **The Horizon 2020 Programme should include a chapter on HLT (as used to be the case until 2015). Innovation projects should be geared at launching the 4-5 powerful domain specific HLT platforms described below.**
        - **Once the European Language Infrastructure and HLT platforms are in place, additional innovation projects should aim at speeding up the delivery of innovative HLT products and services of European and global scale (emphasis added).**
        - **Language interoperability will depend on widespread sharing of EU "big language data." US firms have been able to capitalise on access massive data resources through their platforms to build algorithms. The EC has initiated a European Language Resources Coordination action to gather language data in**

support of the European public services developed in the framework of the CEF Programme. The EC should broaden this effort and make all multilingual data resources created and gathered with public funding available for public AND commercial purposes as a major social and business asset.

  - **It is vital to adjust European copyright law to enable the use of very large data banks in multiple languages for technical purposes such as the training of machine translation engines, by for example “anonymising” data sources.**
  - **Synergies with national initiatives should be sought. In particular, the Member States should become actively involved in ensuring that all their languages are comprehensively taken into account by the ELI.**
  - **Policies in favour of SMEs should aim at removing the “glass ceiling” encountered by scale-up companies in their cross-border growth. Additional efforts are necessary to mobilise VC funding of European scale. A insurance/guarantee scheme for banks funding the international expansion of SMEs would be welcome.**
  - **Mid-tier firms (potential future “blue-chips”) should be included in strategic policy schemes. They are best placed to address high-potential markets that are emerging now and have a key role to play to put Europe at the forefront of innovation.**
  - **The one-sided policy emphasis on open source software should be toned down as it has negative consequences on commercial software product development in Europe.**
  - **Procurement should be used much more intensively to drive innovation in Europe rather than, often ill-targeted and too widespread, subsidy schemes.**

5. **RAISING AWARENESS POLICIES**.

- What needs to be done to improve the perception, surface the availability and increase the acceptance of HLT?
  - **The European Commission and Parliament need to position language as the main barrier for the Digital Single Market.**
  - **The EU (at all levels) should envisage multilingualism as an asset not a liability.**
  - **The EU should invest heavily in HLT to save the tax payer money.**
- What do you think are the most relevant recipients of raising awareness campaigns: policy makers, public officials, firms, citizens, excluded linguistic communities?
  - **All of these need to play their role. However, there needs to be a catalyst. The European Commission is designated by the European Treaties for this role.**
  - **The HLT sector is mobilised and well organised through a number of representative organisations (LT-Innovate for the HLT industry, META-NET for research, GALA for the language service providers, etc.). All these will be able to relay the awareness raising effort, once the European Commission truly embarks**

on providing the catalysing impulse in building the European Language Infrastructure.

6. **DEPLOYMENT POLICIES**.

- In your opinion, what are the policies that could contribute to speed up the effective deployment of HLT in the EU?

- **Europe needs a basic infrastructure for natural language processing (NLP). All language processing applications (search, mining, writing, speech, translation, etc.) depend on such NLP infrastructures. These are tedious to develop and to maintain, and thus expensive, since they are required for every single language.**

- **The European Language Infrastructure (ELI) should provide the basic functionalities required to process unstructured content. Through Application Programming Interfaces (APIs) it should provide basic language technology 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.**

- **The European Commission should use Horizon 2020 funding to launch of 4-5 powerful domain specific HLT platforms. These demonstrator projects could serve as launch pads ("market places") for ecosystems of multilingual applications and services tailored for the specific needs of verticals (e.g. automotive, finance, pharma, healthcare, tourism, etc.). Once the infrastructure (ELI and domain specific HLT platforms) are in place, additional innovation projects aiming at delivering products and services of European and global scale could boost the entire ecosystem.**

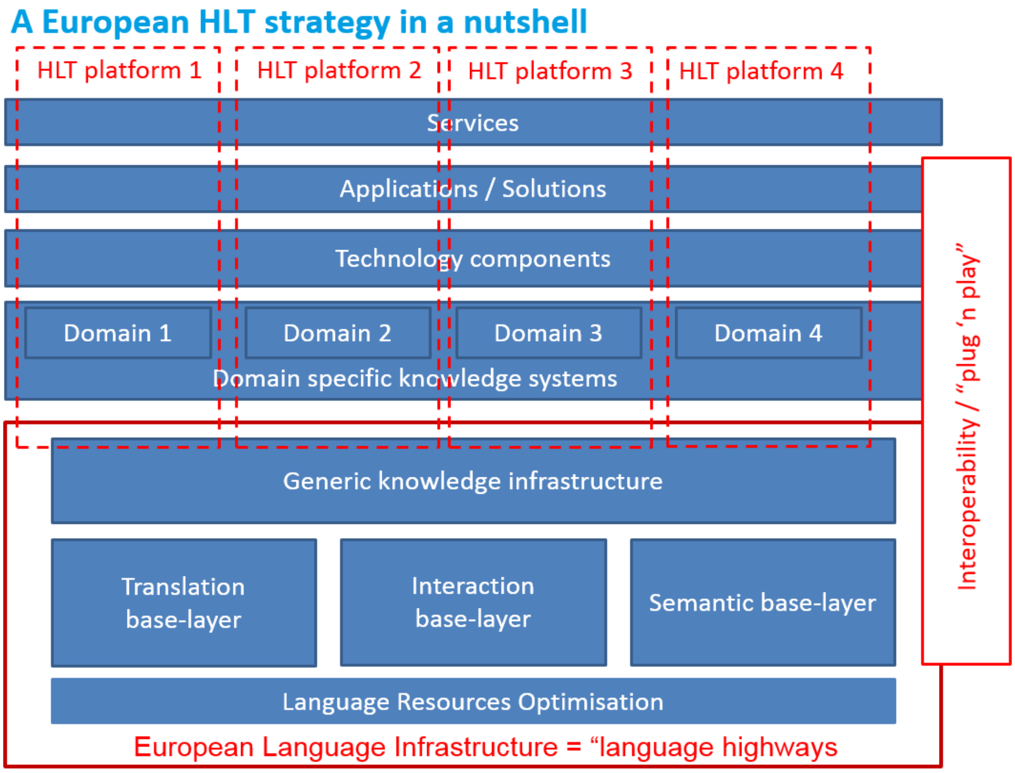
- **Language interoperability, as foreseen in the European Interoperability Framework, is a key concept in achieving this infrastructure, and should operate mandatorily on a par with rail, road, telecoms, information systems, banking systems, and similar interoperable networks that simplify and accelerate traffic and data exchange within the EU and globally. Using public IT standards and shared intelligence, interoperability will mean that both businesses and public agencies will never need to pay constant attention to the "language knowledge" level of their institutions, systems or daily practices. Language interoperability must become as automatic as packet switching around the internet, resulting in a seamless Digital Single Market.**

- **Language interoperability of this kind will involve the use of common APIs at all levels (within the services/applications/technologies/ infrastructure layers and between them) to ensure that meanings are maintained in data between languages expressed in different containers, simplifying the work of all citizens as they search, shop, travel, play, pay and learn.**

- **To set a powerful example of and also benefit from the leverage of interoperability, EU institutions should act as pioneers in this domain, demonstrating its full multilingual potential for citizens needing to consult services in their own languages, and agents wishing to share and exchange information and knowledge of all kinds. The current MT@EC and CEF.AT initiatives constitute a step in the right direction but are, by far, insufficient to achieve the above. Instead, EU Institutions should ensure that the precise**

meaning of information is understood in all languages and that it is preserved throughout exchanges. They can achieve this by maintaining, deploying, and exposing Multilingual Knowledge Systems such as EuroVoc or TMClass.

- **Once a reasonable degree of language interoperability has been achieved by exploiting the instruments suggested in this questionnaire (the ELI and associated HLT platforms) steps will need to be taken to review progress, seek improved solutions, and establish language interoperability as a legal requirement for certain processes throughout the EU. But above all, it should be capable of encouraging business flows, knowledge exchange, innovation and economic growth between countries and their languages.**


A European HLT strategy in a nutshell
HLT platform 1
HLT platform 2
HLT platform 3
HLT platform 4
Services
Applications / Solutions
Technology components
Domain 1
Domain 2
Domain 3
Domain 4
Domain specific knowledge systems
Generic knowledge infrastructure
Translation base-layer
Interaction base-layer
Semantic base-layer
Language Resources Optimisation
European Language Infrastructure = "language highways
Interoperability / "plug 'n play"

  - Create a European platform that makes available all relevant HLT while ensuring that HLT are available as simple, easily usable solutions to speakers of all European languages.

      - **The European Language Infrastructure should NOT make available "all relevant HLT while ensuring that HLT are available as simple, easily usable solutions"! It should concentrate on providing the basic functionalities required to process unstructured content. Through Application Programming Interfaces (APIs) it should provide basic language technology 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.**

    - What do you think would be the best approach to run this platform: a public-private partnership, an academic association, a business association, a heterogeneous consortium, the European Commission or a mix of these different stakeholder groups?

    - **Initially, the European Commission should be the owner of the European Language Infrastructure.**
    - **However, it should procure it from industry (with the support of research where necessary).**
    - **Once the infrastructure is up and running, the European Commission may entrust its governance to a public-private partnership.**

  - Should the platform push one Google Translate-like generic Machine Translation system or many domain/application-specific systems?
    - **The European Commission should not compete with private suppliers. It should merely provide a European Language Infrastructure on top of which generic and domain/application-specific HLT systems can be deployed.**
    - **Supporting the development of domain/application-specific HLT (rather than merely MT) platforms should be one of the objectives of future H2020 innovation projects as the emergence of such platforms is crucial for the competitiveness of European companies.**

- Foster the consolidation of currently fragmented HLT European sector.
  - **Supporting the development of domain/application-specific HLT platforms will have the effect of clustering an "ecosystem" around such platforms and of bringing together the all key players around well-defined value-chains.**

- Foster the growth of small start-up HLT providers through public or mixed investment instruments and accelerator programs.
  - **Yes, start-ups are important. But it is even more important to provide a growth path for scale-up SMEs and mid-tier firms (mature national players who want to Europeanise and/or globalise).**
  - **Many start-ups hit a glass-ceiling (i.e. stop growing at a certain size). This ceiling is due to many structural impediments (first and foremost, the absence of Single Market) which the European Commission should much more proactively help removing / overcoming.**
  - **A healthy "ecosystem" would be composed of start-ups, scale-ups and their corporate and institutional buyers. If, in addition, the research effort could be partly redirected to support this ecosystem through "research-on-demand", well-functioning value-chains would emerge.**

- Facilitate the transfer of technology from language technology providers (academic institutions, research centres, SME companies, large enterprises) to end users (especially SME companies, large enterprises, public bodies etc.). For instance fostering the creation of start-ups providing simple and effective out-of-the-box solutions, particularly in the minority languages markets.

    - **Innovation projects (whether funded by the EC or not) should ALWAYS be driven (i.e. coordinated) by industrial players with research appearing in a support capacity. Technology transfer will be eased once industry “pulls” for research rather than research “pushing” for it.**
    - **Technology providers are already closely involved with end users who are their market, their clients. The EC could add “oil into the cogs” by providing better pan-European market research at affordable prices.**
    - **The minority languages markets will be all the better served when the European Language Infrastructure is available and technologies become cheaper (because they do not need to re-invent the wheel in constantly recompiling the underlying infrastructure elements).**
  - Define specific policies for smaller language communities to enable them to develop and adapt HLT for themselves.
    - **It is very doubtful whether smaller language communities need “specific” policies. Once the European Language Infrastructure is available, all languages will be able to plug into it. National and regional authorities should then support this effort for their respective languages.**
    - **The EU should – as a matter of policy – treat all languages equally. While the implementation schedule of all languages in the European Language Infrastructure may vary due to usage, no language should be overlooked.**
    - **The European Language Infrastructure should also provide a basic infrastructure for the languages of Europe’s main trading partners and/or political neighbours.**
  - Foster freely available open source software, commercial software or a mix of free and commercial software.
    - **This is the remit of technology providers. There is no lack of technologies and/or solutions. The missing links are the European Language Infrastructure, a way for companies to scale-up (overcoming market fragmentation), easy access to investment and a friendly business climate. Rather than substituting itself to commercial providers, the European Commission should focus on making available the latter.**

7. **PUBLIC SERVICE POLICIES**.

- How do you think that the public sector can benefit of HLT to provide multilingual services while contributing to accelerate the development of effective solutions of HLT?
  - Providing effective multi-lingual public e-services while facilitating the access to public information (documents, regulation, open data) to all European citizens regardless of their languages at the European, national, regional and local levels. The EU should co-finance these projects through specific investment funds.

        - **Once the European Language Infrastructure is in place, public services as well as private services will be able to draw upon it to deliver multilingual offerings.**

    - Using 'public procurement of innovative technology' and 'pre-commercial public procurement' to foster the development and deployment of large-scale language technology services and products.

        - **Solutions should be procured wherever possible, as this will have a positive effect on the supply side.**

N.B. The present paper is based on LT-Innovate’s response to a questionnaire issued in November 2016 by the European Parliament's Science and Technology Options Assessment office (STOA) on market and economic impact of the human language technology sector.