Strategic Research and Innovation Agenda

Language as a Data Type and Key Challenge for Big Data

Enabling the Multilingual Digital Single Market through technologies for translating, analysing, processing and curating natural language content

SRIA Editorial Team

Version 0.9 – July 2016

Cracking the Language Barrier

Cracking the Language Barrier

A Federation of European Projects and Organisations working on Technologies for a Multilingual Europe

http://www.cracking-the-language-barrier.eu

The Cracking the Language Barrier federation assembles many European research and innovation projects as well as all related community organisations working on or with cross-lingual or multilingual technologies, in neighbouring areas or on closely related topics. In this umbrella initiative we collaborate on our joint objective to overcome any kind of language and communication barriers with the help of sophisticated language technologies.

Organisations

Projects

This document was prepared by the Cracking the Language Barrier federation. It represents the current state of discussion within the language technology research, development and innovation community towards developing a full Strategic Research and Innovation Agenda for the Multilingual Digital Single Market.

The preparation of this document has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements No. 645357 (CRACKER) and No. 644583 (LT_Observatory).

Executive Summary

The integration of the unified and connected Digital Single Market must address our languages: *The Digital Single Market is a multilingual challenge!* Our treasured multilingualism, one of the cultural cornerstones of Europe and what it means to be and to feel European, is also one of the main obstacles of a truly connected, language-crossing Digital Single Market. The European Language Technology community – including research, development, innovation and other relevant stakeholders – is committed to provide the technologies to achieve this goal.

We recommend setting up the highly focused three-year **Multilingual Value Programme (MLV)** to enable the Multilingual Digital Single Market. This focused programme will be guided by a comprehensive roadmap. It requires a small and modest investment, which can be realised through the Horizon 2020 ICT LEIT funding programme (2018-2020), in close collaboration with the Big Data Value Association (BDVA), i.e., the Big Data Value cPPP.

The **MLV Programme** consists of three application areas that relate to the three main pillars of the Multilingual Digital Single Market. (1) The area **Multilingual E-Commerce** provides multilingual and cross-lingual technologies around search, customer-relationship management, helpdesks, processes, workflows, product catalogues and descriptions etc. (2) The area **Multilingual Content and Media** assembles multilingual and crosslingual technologies for content analytics, curation and generation including authoring support, multimodal and social media. (3) The area **Translation, Language, Knowledge, Data** provides multilingual and crosslingual applications that connect Big Data technologies and Language as well as Knowledge Technologies including machine translation (written, spoken, automatic/human), text mining, business intelligence, sentiment analysis, domain-specific approaches and semantification. These applications are driven by several **Multilingual Services**, which are, in turn, fostered and further improved through **Research**. We also plan to intensify work on basic technologies so that we can cover all relevant languages. In addition, horizontal topics need to be addressed, e.g., standardisation, interoperability, and policy aspects.

The MLV Programme will not only unlock the multilingual Digital Single Market through a set of platforms, services and solutions that support all businesses and citizens, it will provide the European language technology community and several different industries with the ability to compete with other markets and achieve multiple benefits for the European economy and future growth, as well as for society and the citizens. To achieve this ambitious plan, all stakeholders need to collaborate and cooperate closely and in a tightly coordinated way. To demonstrate that the whole Multilingual Europe community firmly stands together, this document is presented by the Cracking the Language Barrier federation, which consists of 10 organisations and more than 20 projects working together on the technological foundations of a Multilingual Europe.

Awareness, political determination and will are required to make sure that the Digital Single Market takes the language component into account. VP Andrus Ansip's recent blog post, "How multilingual is Europe's Digital Single Market?" is a sign that the awareness is there – now it is simply a matter of making sure that the MLV Programme can be put into practice.

By realising the Multilingual Digital Single Market, the MLV Programme would solve the issue of language-blocking and language-induced market fragmentation. It would also reduce the threat of digital language extinction. We recommend that Europe actively makes an effort to compete in the global landscape for research and development in language technology since we cannot expect third parties from other continents to solve our translation and knowledge management problems in a way that suits our specific communicative, societal and cultural needs.

Language Technology made for Europe in Europe is the key. It will contribute to future European cross-border and cross-language communication, economic growth and social stability.

Table of Contents

1. The Multilingual Value Programme (MLV) for the Multilingual Digital Single Market

The Digital Single Market (DSM) holds tremendous potential to transform the European economy and make it more globally competitive. However, *one single* digital European market as such does not yet exist: it is still a collection of many separate smaller markets, confined by national or regional language boundaries. By contrast, China or the United States represent truly national markets. It is no surprise that most of the pioneering growth in ecommerce has happened in the US, where regulatory barriers are lower and a single language can address the vast majority of the market. Europe needs to open up the invisible borders created by our different languages. All of the languages actively spoken in Europe are also used digitally: ecommerce shops, information pages, online services, encyclopedias, university pages, company websites, user-generated content, online videos, podcasts, radio stations, and other multimedia content all make use of the official, regional, and unofficial minority languages spoken in Europe. These languages must also be covered and reflected by the Digital Single Market. To realise this, we suggest to put into place applications, platforms and services based on language technologies.

We recommend setting up the highly focused three-year **Multilingual Value Programme** (**MLV**) to enable the Multilingual Digital Single Market. The MLV Programme will be guided by a comprehensive roadmap. It requires a small investment, which can be realised through the Horizon 2020 ICT LEIT funding programme (2018-2020), in close collaboration with the Big Data Value Association (BDVA), i.e., the Big Data Value cPPP. The MLV Programme consists of three multilingual application areas that relate to the three main pillars of the Multilingual Digital Single Market. (1) Multilingual E-Commerce; (2) Multilingual Content and Media; (3) Translation, Language, Knowledge and Data Applications. These applications are driven by several **Multilingual Services**, which are, in turn, fostered and further improved through **Research**. We also plan to intensify work on basic technologies so that we can cover all relevant languages. In addition, horizontal topics need to be addressed, e.g., standardisation, interoperability, and policy aspects. The following Chapter will discuss the programme in more detail.

The European Commission predicts that the transition to the integrated DSM will deliver up to €400 billion in economic growth by 2020. Measures like eliminating roaming charges, improving legislation (especially copyright and data protection), and making cross-border payments easier are all important and necessary preconditions. However, they are not sufficient to accomplish the overall goal. If customers are hampered by language, online commerce will remain confined to fragmented markets, defined and restricted by language silos. Even the unacceptable suggestion for everyone to use English would not deliver a single market, since less than 50% of the EU's population speaks English, and less than 10% of non-native speakers are proficient enough to use English for online commerce. Approximately 60% of individuals in non-Anglophone countries seldom or never make online purchases from English-language sites; the number willing to purchase from sites in non-native languages other than English is much, much lower.[1]

As a result, no single language can address 20% or more of the DSM (German comes closest, as the native language of 19% of the EU's population). Taking care of the top four EU languages (German, French, Italian, English) would still address only half of the EU citizens in their native language. Even allowing for second-language speakers, no single language can address more than a fraction of the DSM. Concentrating exclusively on the 24 official EU languages would exclude those European citizens from the DSM who speak regional or minority languages, languages of important trade partners or languages of refugees.

[1] Common Sense Advisory (2014): "Survey of 3,000 Online Shoppers Across 10 Countries Finds that 60% Rarely or Never Buy from English-only Websites", https://www.commonsenseadvisory.com/Default.aspx?Contenttype=Article Det&tabID=64&moduleId=392&Aid=21500.

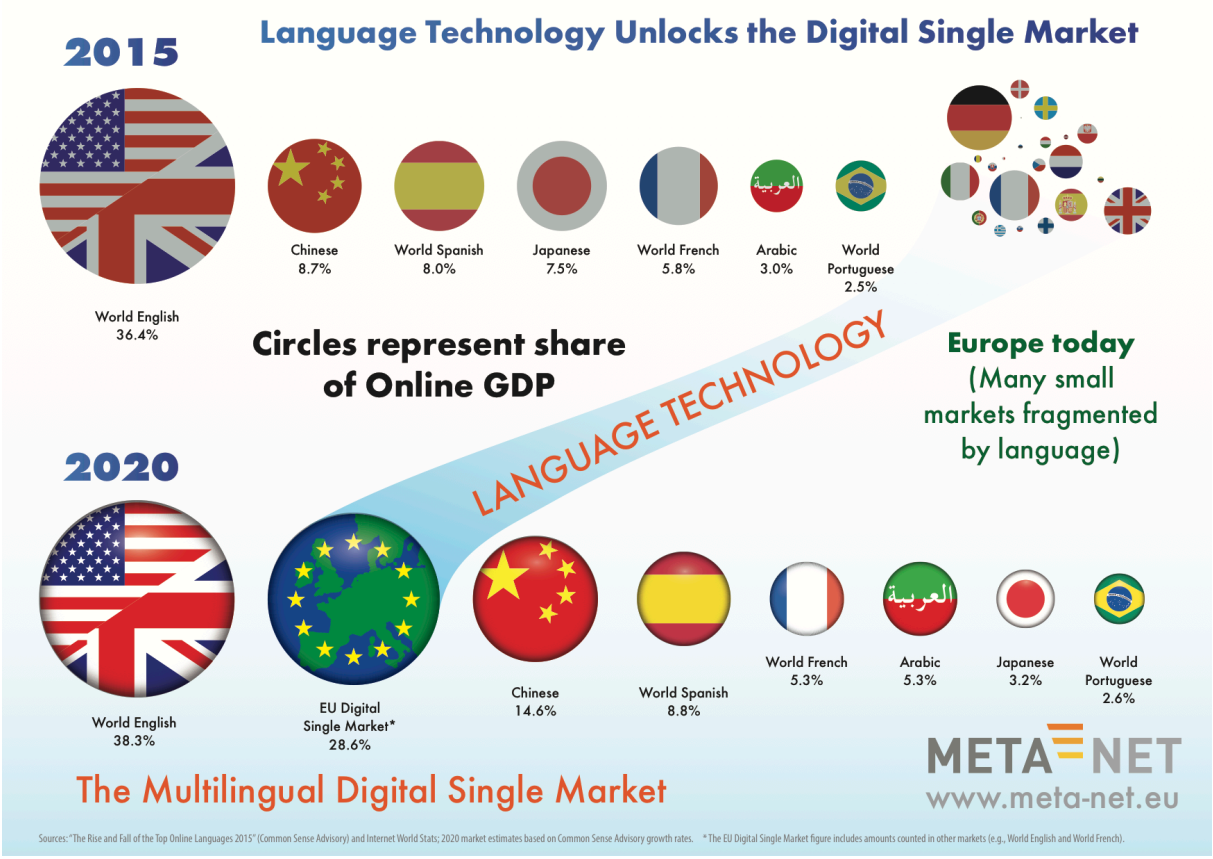


Figure 1: *Language technology unlocks the Digital Single Market*

Small and medium-sized European companies are a vital component of the DSM. However, only 15% of European SMEs sell online – and of that 15%, fewer than half do so across borders.[2] SMEs that sell their products and services internationally exhibit 7% job growth and 26% innovate in their offering – compared to a job growth of 1% and 8% innovation for SMEs that do not sell their products and services internationally.[3] Only if Europe accepts the multilingual challenge and decides to design and to implement research and innovation-driven technology solutions as well as a service infrastructure with the goal of overcoming language barriers, can the full economic benefits of the DSM be achieved. Enabling and empowering European SMEs to easily use language technologies to grow their business online across many languages is key to boosting their levels of innovation and to help them create jobs.

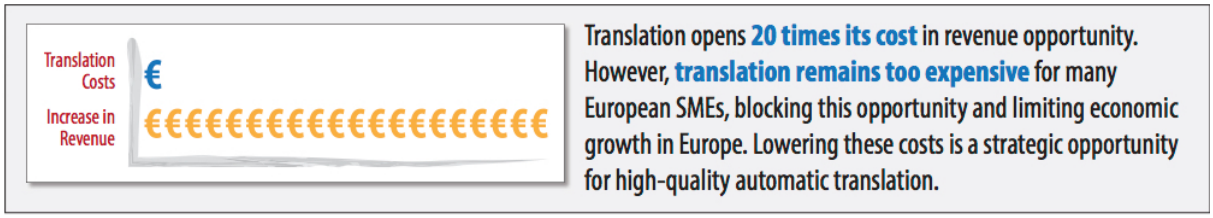


Figure 2: *Translation opens 20 times its cost in revenue opportunity*

The European DSM today would account for approximately 25% of global economic potential. However, if Europe were to overcome the language barriers that hamper intra-European trading, it would also remove barriers to international trade that keep European SMEs from achieving their full economic potential by entering and penetrating markets in other continents beyond our own. Addressing the official and major regional languages of Europe would open access to over 50% of the world's online potential and 73% of the world online market in economic terms, amounting to an

[2] EC (2015): "How digital is your country? New figures reveal progress needed towards a digital Europe", http://europa.eu/rapid/press-release_IP-15-4475_en.htm.

[3] EUbusiness: "Annual Report on European SMEs 2013-14 – A Partial and Fragile Recovery", http://www.eubusiness.com/topics/sme/report-2014.

online market of approximately €25 trillion (sic) (in 2013).[4] Most of this increase comes from English, Spanish, French, and Portuguese, but other languages also make significant contributions to world-wide market access. The *global potential* for European businesses exceeds the *continent-internal* opportunities from the DSM by orders of magnitude.


European Commission › ··· ›

BLOG POST | By Andrus Ansip | 27 May 2016

How multilingual is Europe's Digital Single Market?

Europe's vast marketplace is unique in the world for several reasons. This is not merely because of its sheer size – more than 500 million consumers – but also because of its impressive cultural and linguistic diversity.

While we may not think of it very often, the European Union is home to 24 official languages: Slovenian and Swedish, French and Finnish, Polish and Portuguese. And that is only a quarter of them.

Figure 3: *Blog post by Andrus Ansip, published on 27 May 2016*

At the end of May 2016, VP Andrus Ansip published a blog post titled "How multilingual is Europe's Digital Single Market?".[5] In his article, VP Ansip not only acknowledges that Europe's multilingualism "brings difficulties for people and businesses to understand each other and to operate across borders". Especially in ecommerce the language barrier can be a concrete obstacle, VP Ansip uses the very adequate phrase "don't understand, won't buy" to describe the situation that especially affects smaller online retailers and web-based traders. As a consequence, online shops often provide (at least) 24 different language versions of their website but it does not end with the actual sale of an item: after-sales services with the same multitude of languages also need to be taken into account. A similar situation exists in the area of data analytics, where data sets in different languages cannot be easily aggregated or semantically processed. VP Ansip mentions that the previous investments of the EC in language technology-related projects (including infrastructural services such as CEF Automated Translation), recent advances in Machine Translation and other multilingual technologies have the potential of breaking through language barriers. Andrus Ansip's goal is "to turn Europe's linguistic diversity from a barrier into an asset" since the DSM "is by definition multilingual".

This Strategic Agenda and Roadmap is meant to be the next step with regard to VP Ansip's goal of reducing and finally removing the language barriers that are holding back the advance of the Digital Single Market and to turn them into competitive advantages.

1.1. Overcoming Language Barriers with Language Technologies

The borders between our languages are invisible barriers at least as strong in their separating power as any remaining regulatory boundaries. They create fragmented and isolated digital markets with no bridges to other languages, thereby hampering the free flow of products, commerce, communication, ideas, help, and thought. Language barriers in the online world can only be overcome by (1) significantly improving one's own skills in non-native languages, (2) making use of others' language skills, or (3) through digital technologies. With the 24 official EU languages and dozens of additional languages, relying on the first two options alone is neither realistic nor feasible. For specific types of

[4] Benjamin B. Sargent, Common Sense Advisory (2013): "The 116 Most Economically Active Languages Online", https://www.commonsenseadvisory.com/AbstractView.aspx?ArticleID=5590.

[5] https://ec.europa.eu/commission/2014-2019/ansip/blog/how-multilingual-europes-digital-single-market_en.

content and purposes, specialised human language services, increasingly assisted by language technology themselves, will continue to play a major role in translating documents, creating subtitles for videos, or localising websites into 20+ other languages. However, relying on human services alone would exclude most SMEs from the DSM because of the high costs. It would create a market that can only be successfully penetrated by large, consolidated enterprises, which is why cost-effective methods must be found to support market access for SMEs and European citizens.

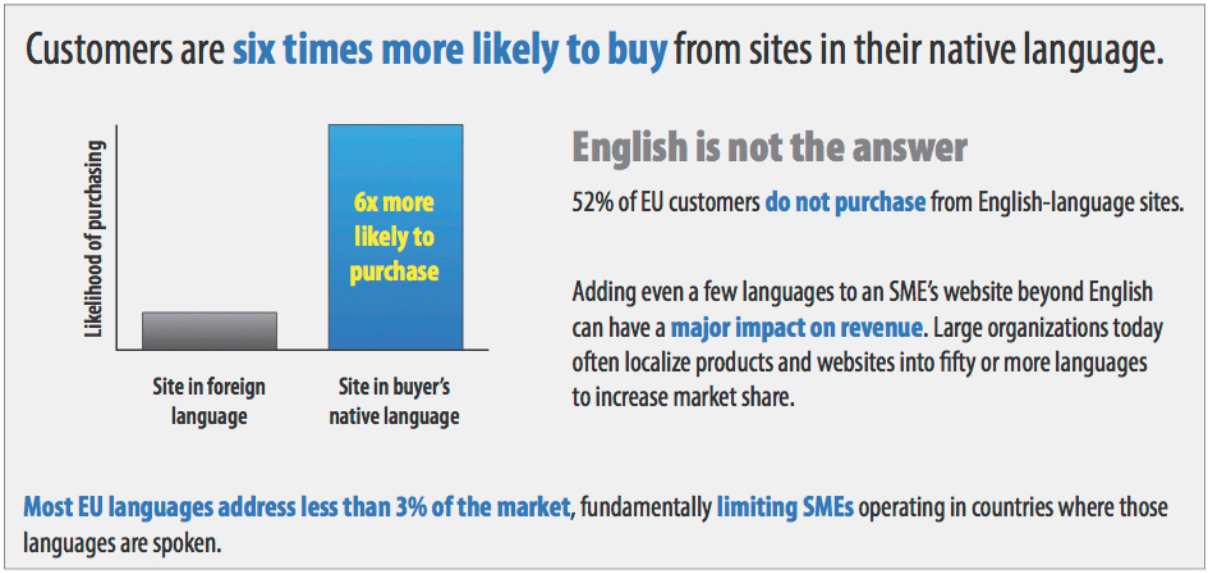


Figure 4: *The impact of presenting an online shop in the customer's native language*

To succeed, any SME must both excel in communicating its expertise in its market niche and be able to engage in two-way conversations with its customers online. The free machine translation services offered by a few US-tech giants are useful for giving users the gist of web content. But they cannot be easily and cheaply tailored to support the niche communication needs between SMEs and their customers. Supplementing this with domain-tailored language services such as, for example, content and sentiment analysis, knowledge extraction and multimodal online engagement is completely out of reach for SMEs aiming to engage the half of the EU consumers who do not enjoy English, German, French or Italian as their native language.

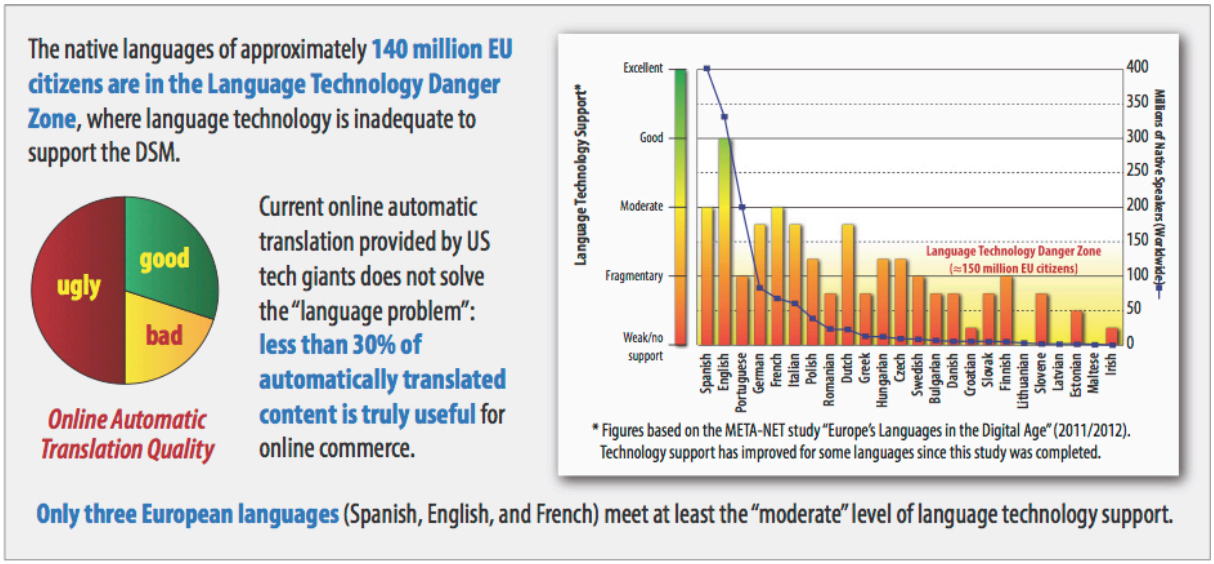


Figure 5: *Many European languages only have very weak technology support*

The connected and truly integrated Digital Single Market can only exist once all language barriers have been overcome and all languages are connected through technologies. Only advanced communication and information technologies that are able to process and to translate spoken and written language in a

fast, robust, reliable, and ubiquitous way, producing high-quality output, can be a viable long-term solution for overcoming language barriers.

Establishing such an infrastructure requires a big collective push that involves designing, implementing and deploying technologies, services and platforms, accelerating innovation, research and efficient technology transfer. While only a few of our languages are in a moderate to good state with regard to technology support, more than 70% of our languages are seriously under-resourced, actually facing the danger of digital extinction (for example, Maltese, and Lithuanian), even though it must be noted that support for these languages with smaller numbers of speakers is slowly increasing (cf. the Appendix).[6]

Today's IT systems are only just beginning to handle the meaning, purpose and sentiment behind our trillions of written and spoken words. Language makes up a very large part of the continuously growing Big Data treasure. Today's computers cannot understand texts and questions well enough to provide high-quality translations, precise summaries or reliable answers in all languages. Yet in less than ten years such services could be offered for many. Technological mastery of human language can enable a multitude of innovative IT products and services in industry, commerce, government and administration, private and public services, education, healthcare, entertainment, tourism and many other sectors.

Language technology is therefore the missing piece of the puzzle that will bring us closer to a fully integrated DSM. But language technology does more than enabling the DSM. It is a key technology for the next generation IT, which will be much smarter and human-centered in its functionality. Almost every digital product uses and is dependent on language – which is why language technology is a mandatory component! It is the key enabler to boosting growth in Europe and strengthening our competitiveness in a sector that has become critical for Europe's future, considering the significance given to the DSM by the EU.

The European countries and language communities constitute a set of individual, unconnected, fragmented, isolated markets. A truly integrated DSM that spans our continent can never exist if we ignore the "language factor" and the de facto state of play: European citizens are unable to access vast amounts of online content due to *language-blocking*. The European economy is suffering as well because there are no technical means that enable, say, a restaurant owner in Latvia to order ten crates of wine in Portugal if the restaurant owner, who speaks Latvian, is unable to find the website of the vineyard, presented in Portuguese, in the first place. And negotiating and completing a deal would require a translator.

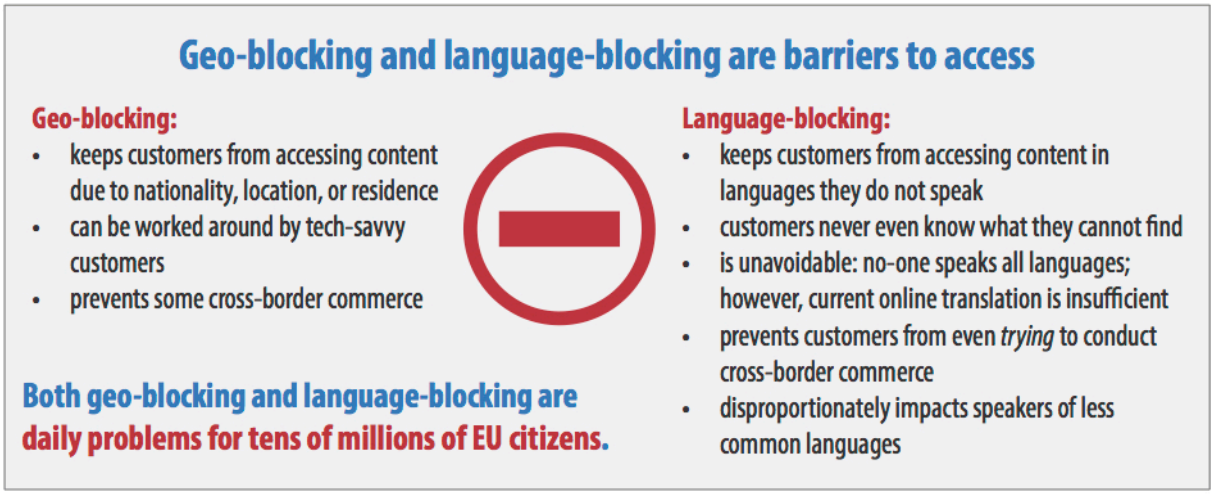


Figure 6: *Language-blocking is a barrier to access, just as geo-blocking is*

Europe is the most appropriate place for realising the MLV Programme by virtue of applied research, development and innovation. Our continent has half a billion citizens who speak one of over 60

[6] See the results of the META-NET White Paper Series, http://www.meta-net.eu/whitepapers.

European and many non-European languages as their mother tongue. Europe has more than 2,500 small and medium-sized companies working on language, knowledge and interface technologies, and more than 5,000 companies providing language services that can be improved and extended by technology. Europe has a long-standing research, development and innovation tradition with over 800 centres performing excellent, highly visible and internationally recognised research on all European and many non-European languages.

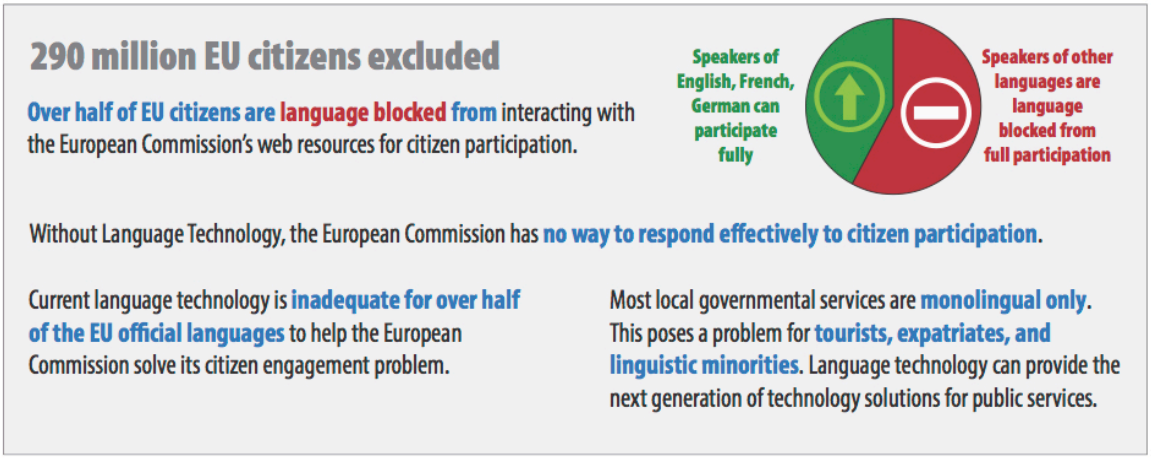


Figure 7: *Selected effects of language-blocking*

1.2. Language as a Data Type – Multilingual Big Data Content Analytics and Generation for the European Data Economy

Language is not only a necessary ingredient of the DSM, it is a mandatory enabler for the future European Data Economy. Data is the oil of the 21st century. Data linking and content analytics are key technologies for refining this oil so that it can drive the engines of understanding – data homogenisation, semantic analysis, enrichment, repurposing. Large data sets are never solely numerical data – they always come with language components such as column heads in database tables, free text in table cells, metadata annotations, descriptions, documentation, summaries, links to specific documents etc. The Data Economy requires innovative new mechanisms that enable data and data value chains to flow freely across language boundaries.

Figure 8: *Multilingual data value chains*

We also need to pay attention to the sheer volume of data generated. Only one hour of customer transaction data at Wal-Mart, corresponding to 2.5 petabytes of data, is 167 times the amount of data housed by the Library of Congress.[7] Data growth keeps rising: 90% of the data available today has been generated in the past two years only.[8] IDC estimates that all digital data created, replicated or consumed will grow by a factor of 30 between 2005 and 2020, doubling every two years. By 2020, it is assumed that there will be over 40 trillion gigabytes of digital data, corresponding to 5,200 gigabytes per person on earth.[9] The Internet of Things will add not only more but additional types of data (including large amounts of textual data, of course): Cisco estimates that currently less than 1% of physical objects are connected to computer networks. According to recent estimates by Cisco this number will rise to up to 50 billion connected devices by 2020, corresponding to between 6 and 7 devices per person on the planet. Europe needs a scalable technological infrastructure for handling its big data sets. While the specific Big Data solutions circling around computer science and advanced database technologies will be taken care of by the Big Data Value Contractual Private Partnership (BDV cPPP), the examples given above demonstrate the need for complementary language technologies and for creating synergies between the BDV cPPP and the European Language Technology community by including robust and precise multilingual text analytics technologies that can perform at web-scale level and, even more crucial, at an Internet of Things level.

Figure 9: *The Strategic Agenda of the Big Data Value Association*

Big Data analytics will not just be "slightly better" if we include language technology – it simply will not happen without language technology. We cannot simply put any type of Big Data into a database and then build applications on top of it – we will need to process it sensibly and that sense will need to be based on language. This challenge not only relates to structured Big Data, which itself typically exists only in a language silo, but especially to any type of *unstructured* data (i.e., Linguistic Big Data) including text documents and social media streams, essentially any sequential symbolic process of meaningful information. Language technologies will build bridges from Big Data to Knowledge, from Unstructured Data to Structured Data. Language Technology will become the foundation for organising, analysing and extracting data in a truly useful way, it must be and will become a necessary ingredient in any monolingual or cross-lingual data value chain.

The MLV Programme can only be successful if we engage in a close, complementary collaboration with the BDV cPPP to ensure that the multilingual and cross-lingual Big Data value chains reflect the subtleties and variety of language in the use of vocabulary, register, idioms and tone that is distinct to individuals, communities and domains – in that sense, the present Strategic Agenda and Roadmap provides further insights with regard to treating "Language as a Data Type" in Big Data Applications. The European Big Data Value Strategic Research and Innovation Agenda already mentions the need for complementary research and innovation activities on Linguistic Big Data: "In Europe, text-based data resources occur in many different languages, since customers and citizens create content in their local language. This multilingualism of data sources makes it often impossible to use existing tools and to align available resources, because they are generally provided only in the English language. Thus, the seamless aligning of data sources for data analysis or business intelligence applications is hindered by the lack of language support and availability of appropriate resources." (p. 23).[10]

---

[7] Beñat Bilbao-Osorio et al. (ed.) (2014): "The Global Information Technology Report 2014 – Rewards and Risks of Big Data", World Economic Forum and INSEAD.

[8] SINTEF (2013): "Big Data, for better or worse: 90% of world's data generated over last two years".

[9] John Gantz and David Reinsel (2012): "The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East", International Data Corporation (IDC).

[10] BDVA (Big Data Value Association): "Big Data Value Strategic Research and Innovation Agenda" (V2.0), http://www.bdva.eu/sites/default/files/EuropeanBigDataValuePartnership_SRIA__v2.pdf (2016).

The overall strategic goal connected to the BDVA SRIA is to "deliver new Big Data technology allowing for deep analytics capacities on data-at-rest and data-in-motion while providing sufficient privacy guarantees, optimized user experience support and a sound data engineering framework." The BDVA SRIA specifies five technical priorities. Of these, Data Management, Data Processing and Data Analytics are especially relevant to Language Technologies:

- **BDVA SRIA Technical Priority "Data Management":** The specific connection to the language topic and main challenge is data in a variety of formats including spoken data and text data in multiple different languages. Needed outcomes are techniques and tools for handling unstructured and semi-structured data including NLP methods for different languages, algorithms for the automatic detection of abnormal structures (e.g., social, geospatial and other domain oriented data), as well as standardised annotation frameworks for different sectors supporting the integration of annotation technologies and data formats. Furthermore, techniques for the semantic interoperability such as standardised data models and interoperable architectures for different sectors are needed including standards and multilingual knowledge repositories that allow the seamless linking of data.
- **BDVA SRIA Technical Priority "Data Analytics":** For this technical priority, outcomes in terms of improved, more accurate statistical models are needed, especially with regard to semantic analysis. Deep learning, contextualisation, machine learning, NLP, smart data analytics and real-time semantic analysis are on the agenda, including event and pattern discovery (natural disasters, rare diseases etc.). This technical priority also includes methods for unstructured multimedia analytics and data mining, linking and cross-analysis algorithms to deliver cross-domain and cross-sector intelligence.
- **BDVA SRIA Technical Priority "Data Processing":** The main language-related challenge for this priority is real-time analytics and event processing, especially of highly heterogeneous data sources and data formats that need to be processed together and linked as well as aligned with one another, including highly heterogeneous semantic representations, various levels of granularity, unstructured, semi-structured and structured data, text and multimedia data etc. A key goal is to extract knowledge out of these heterogeneous data sets, which puts special emphasis on the quality, precision and robustness of the respective language technologies.

In the further specification of Chapter 2, we put special emphasis on combining and addressing synergies between the needs of the Multilingual Digital Single Market and the needs and requirements expressed within the BDVA SRIA document.

1.3. Increasing Demand for Multilingual Content

There is increasing pressure to overcome language barriers: online content in hitherto dominant languages is declining and "long-tail" languages are rising.[11] In line with the constant rise of content, absolute numbers are rising for all languages, and much more significantly so for less common languages. For example, Basque, Galician, and Catalan all have an increasing share vis-a-vis Spanish; even though the numbers are small, they indicate a long-term shift.

This trend goes hand in hand with increasing public demand for content in regional or local languages due to the increasing availability of broadband as well as mobile connectivity and rising numbers of online users and online services. Europe is no longer satisfied with using only a few major languages. As a consequence, businesses that cannot provide content in local languages will be global losers. Market saturation for dominant languages has been reached, any additional growth is coming from outside the established markets, historically served by a smaller set of languages.[12] If we extrapolate the trends reported by Common Sense Advisory, it only took 37 languages to reach 98% of the world online population in 2009, but already 48 in 2012. The predicted number in 2015 is 62 languages.

[11] Common Sense Advisory (2013): "The Rise of Long-Tail Languages".

[12] ibid.: "Traditional "power house" languages are seeing some of the biggest drops in overall site support: e.g., German: -11.7%, French: -13.4%, Spanish -14.4%, i.e., a smaller percentage of "global" sites are supporting these languages, even as the number supporting long-tail languages is increasing".

More and more citizens are connected and, as a consequence, more and more citizens use – and expect to use – their native languages in online activities. However, they are often excluded from participating due to the fact that language barriers constitute market barriers – especially so with regard to the DSM. True engagement with consumers across language barriers is also deeply entwined with the user's technical, cultural and individual awareness, preferences and requirements. The power of personalising any cross-linguistic exchange to an individual user means we should not merely bridge the language barrier but provide a compelling and personalised user experience.

The impact a truly connected DSM could have is not just felt in terms of sales. Technological integration fails if content cannot be used. For example, electronic standards for integrating health records simply add cost without benefit if the recipient is not able to interpret and use those records. If doctors' notes and observations remain in one language and are not accessible, they cannot help doctors other regions, e.g., if a traveller from Poland falls ill while in France. Here the impact of language barriers is measured not just in terms of Euros but in terms of health and, potentially, lives.

1.4. EC and Language Technology – Past and Present

Already in the late 1970s the EU realised the relevance of language technology as a driver of European unity and began funding its first research projects, such as EUROTRA (1978-1992). After a longer period of sparse funding, the EC set up a department dedicated to language technology and machine translation; it was later integrated into the new "Data Value Chain" unit in DG Connect (Directorate General for Communications Networks, Content and Technology).

In recent years, the EU has been supporting projects such as EuroMatrix, EuroMatrixPlus (2006-2008, 2009-2012), Let'sMT! (2010-2012), and iTranslate4 (2010-2012), which draw on basic and applied research along with industrial collaboration to generate machine translation resources for many European languages. More recently, the large-scale META-NET initiative (supported in its first phase by four EU projects), which started in 2010, has assembled the Language Technology community around its core network of excellence which consists of 60 research centres in 34 European countries: META, the Multilingual Europe Technology Alliance, has more than 800 members. META-NET has prepared studies such as its 30-volume White Paper Series, and the META-NET Strategic Research Agenda for Multilingual Europe.[13] The open resource exchange infrastructure META-SHARE provides access to thousands of language resources and technologies. The EU has also facilitated the coalescing of the LT industry through the FP7 support action LT COMPASS. The resulting industry association, LT-Innovate, currently counts 180 corporate members. LT-Innovate issued a Report on the State of the European Language Technology Industry[14] and an Innovation Agenda[15]. At the beginning of 2015 new projects have been launched, funded through the Horizon 2020-ICT 17 call. In addition to the large research action QT21, which is working on new paradigms for high-quality machine translation, three innovation actions are adapting and applying new MT methods for industrial and commercial use cases. In the middle of 2015, the EU project CRACKER initiated the "Cracking the Language Barrier" federation of organisations and projects working on technologies for a Multilingual Europe. This umbrella initiative is continuously getting more members and currently assembles 10 organisations and more than 20 projects.

In parallel to the research and innovation-oriented activities funded through FP7 and Horizon 2020, the EC is further advancing the Connecting Europe Facility programme (CEF). Part of CEF Telecom is the Automated Translation building block that "helps European and national public administrations exchange information across language barriers in the EU" and also to make all of CEF's Digital Service Infrastructures multilingual.[16] This Automated Translation service, CEF AT, builds on an existing machine translation system, MT@EC, developed within the EC (DG Translate). It is being

[13] See http://www.meta-net.eu/whitepapers and http://www.meta-net.eu/sra.

[14] LT-Innovate Innovation Agenda & Manifesto (2014): "Unleashing the Promise of the Language Technology Industry for a Language-neutral Digital Single Market".

[15] LT2013 (2013): "Status and Potential of the European Language Technology Markets".

[16] Connecting Europe Facility (CEF): "Automated Translation", https://joinup.ec.europa.eu/community/cef/og_page/ catalogue-building-blocks#AT.

implemented on the Moses toolkit, under the Interoperability Solutions for European Public Administrations (ISA) programme. One of the key ideas is to harness the linguistic knowledge embodied in the EC's database of translated documents covering the 24 official languages of the EU. MT@EC is currently only available to staff members of the EC and the EP as well as public administrations of EU member states. A closer collaboration between CEF AT and the European language technology community has been established through the service contract ELRC (European Language Resource Coordination) which was awarded in September 2015, especially with regard to the systematic and coordinated collection and exploitation of language resources in all CEF participating countries. Follow-up service contracts are expected for late 2016.

Looking beyond the EC, research by TAUS[17] has shown that European research funding that fostered the development of the open source MT toolkit Moses has opened up new business opportunities in language technology by enabling companies to reduce the cost required to translate content, particularly in fields such as technical support. These cost reductions have helped companies to increase their multilingual reach and engage with customers in language markets inaccessible through traditional translation routes. There is a clear long-term trend to increasing language support and increasing customer engagement via language technologies. According to the report, there are already 22 operative Moses-based MT companies with an estimated market share of about $45 million or about 20% of the entire MT solutions market.

1.5. The Economic Power of Language Technology and Services

In addition to being a key enabling technology for the multilingual DSM, Language Technology comes with a non-trivial economic power itself. The European market for translation, interpretation and localisation was estimated to be €5.7 billion in 2008. The subtitling and dubbing sector was at €633 million, language teaching at €1.6 billion. The overall value of the European language industry was estimated at €8.4 billion and expected to grow by 10% per year, i.e., resulting in ca. €16.5 billion in 2015. The global language technology industry[18] was evaluated at €26.5b in 2015, projected to rise to €65b by 2020. Driven through interactive dialogue systems and smart personal assistants such as Apple Siri, Microsoft Cortana, Amazon Echo and Google Home, the global speech technology market alone will reach ca. US$20.9 billion by 2015 and ca. US$31.3 billion by 2017. Yet, this existing capacity is not enough to satisfy current and future needs, e.g., with regard to translation. Today, Google Translate translates the same volume per day as all human translators on the planet translate in one year and is used by more than 200 million people every month.[19]

2. The Multilingual Value Programme (MLV)

The Digital Single Market must address our languages. Our treasured multilingualism, one of the main cultural cornerstones of Europe and what it means to be and to feel European, is also one of the main obstacles of a truly connected *Multilingual Digital Single Market*. The European Language Technology community – including research, development, innovation and other relevant stakeholders – is committed to provide the technologies to achieve this goal.

We recommend setting up the highly focused three-year **Multilingual Value Programme** (**MLV**) to enable the Multilingual Digital Single Market. The MLV Programme will be guided by a comprehensive roadmap. It requires a modest and small investment, which can be realised through the Horizon 2020 ICT LEIT funding programme (2018-2020), in close collaboration with the Big Data Value Association (BDVA), i.e., the Big Data Value cPPP.

[17] Achim Ruopp, Jaap van der Meer, TAUS (2015): "Moses MT Market Report", https://www.taus.net/think-tank/reports/ translate-reports/moses-mt-market-report.

[18] Figures from LT2013: Status and Potential of the European Language Technology Markets, April 2013.

[19] http://googleblog.blogspot.de/2012/04/breaking-down-language-barriersix-years.html.

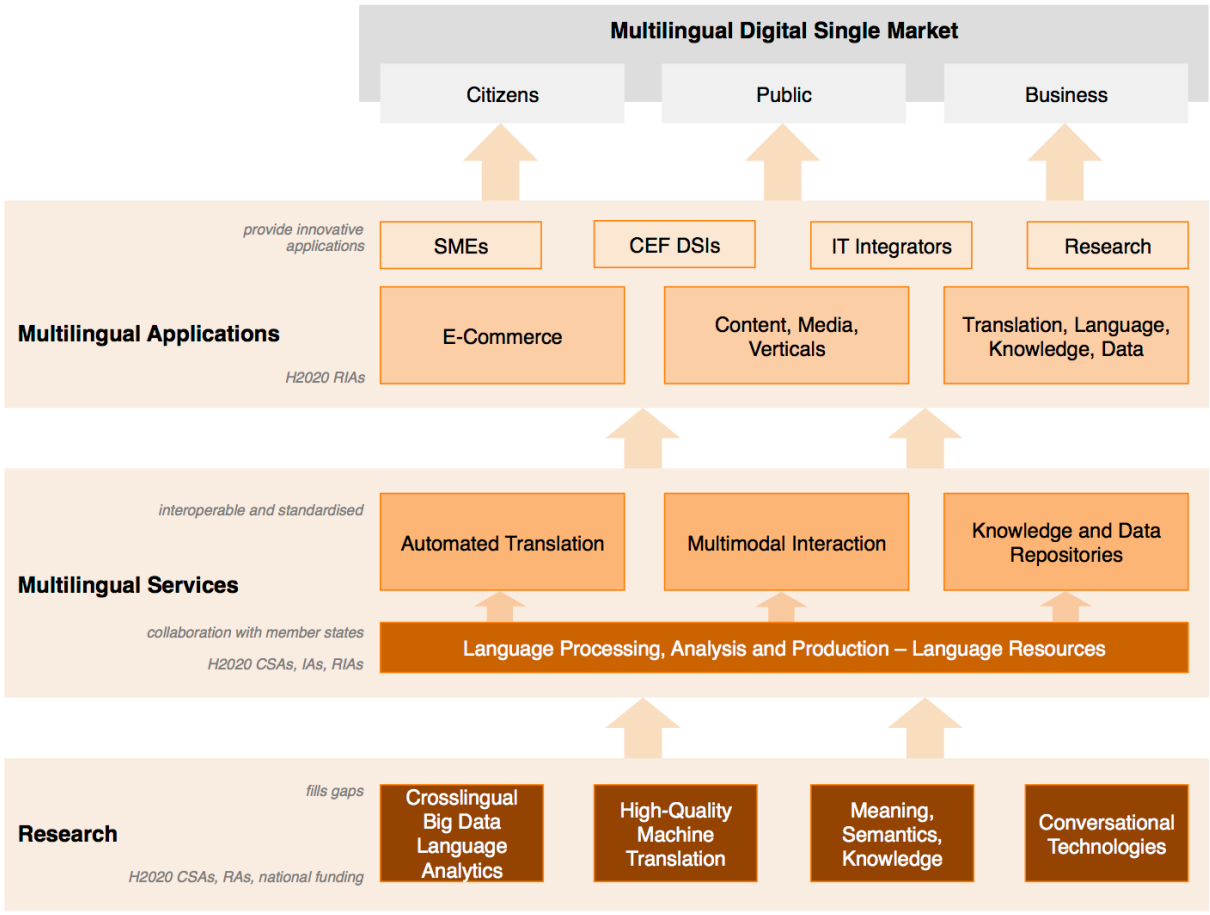


Figure 10: *The Multilingual Value Programme*

The **MLV Programme** consists of three **Application Areas** that relate to the three main pillars of the Multilingual Digital Single Market.

1. The area **Multilingual E-Commerce** provides multilingual and cross-lingual technologies around search, customer-relationship management, helpdesks, processes, workflows, multilingual product catalogues and descriptions etc.
2. The area **Multilingual Content, Media, Verticals** assembles multilingual and crosslingual technologies for content analytics, curation and generation including authoring support, multimodal and social media. It also includes the vertical domains health, legal, government, mobility and energy.
3. The area **Multilingual Language, Knowledge and Data Services** provides – as user- and customer-oriented applications – multilingual and crosslingual services that connect Big Data technologies with Language as well as Knowledge Technologies including machine translation (written, spoken, automatic/human), text mining, business intelligence, sentiment analysis, domain-specific approaches, semantification.

The corresponding applications in these three areas are solutions that use the Multilingual Services available in the second layer (see Figure 10). The applications enable e-commerce and digital service providers, public administrations, cross-border public service providers, and other stakeholders to easily integrate multilingual capabilities in their daily work. The solutions can also be integrated into the workflows of large-scale organisations – such as European institutions, NGOs, media outlets, and corporations – as well as be made available to the public, enabling multilingualism on a pan-European scale.

The Multilingual Applications are driven by a set of lower-level **Multilingual Services**. Among these are services for **Automated Translation**, **Multimodal Interaction** as well as **Knowledge and Data Repositories** (including services to query and to enrich these repositories). The fourth set has a special status, these are **Language Processing, Analysis and Production** services, e.g., typical Natural

Language Processing (NLP) services, and also **Language Resources**, i.e., language data sets of different types. All of these services need to be modular, interoperable and standardised so that they can be freely and seamlessly integrated within each other and also within the set of Multilingual Applications. This is why we suggest that the design and implementation of the multilingual services – as part of innovation actions – feeds into the development of standards in the following selected areas and their respective standardis developing organisations: data quality (including translation and general), multilingual and cross-format content analytics workflows and interfaces, cross-lingual data integration, data and knowledge representation formats. The multilingual services provide fundamental facilities and systems which private companies cannot build efficiently or profitably on their own. While, in general, first versions of these services and applications can be devised and deployed quite rapidly due to previous EC investments, they need to be supported and continuously updated through research. Finally, we suggest to concentrate on the four main research themes mentioned in Figure 10. In addition, we plan activities related to platforms and basic language resources and technologies so that all relevant languages are covered. Several Horizontal Topics need to be addressed, e.g., standardisation (see above) and policy aspects.

The Multilingual Digital Single Market has a large set of very specific needs in terms of technologies for overcoming language borders. The Multilingual Applications, Multilingual Services and Research layers we propose are, with regard to the initial version 0.5 of this strategic agenda, an improved and extended suggestion. They need to be discussed, focused and prioritised with the European Commission and also with all involved stakeholders before we can arrive at the final version 1.0 of this document.

2.1. Technical Approach

The MLV Programme will result in a set of services (in the sense of cloud-services, web-services, RESTful services, application programming interfaces etc.) that drive the Multilingual Applications, sketched in the roadmap (see Chapter 3). They can be conceptualised as Software-as-a-Service, but also as components that can be integrated into stand-alone software. We suggest to start with a small set of clearly defined, mission-critical services needed by the majority of applications. This initial set of seed services, then, needs to be able to scale organically into one or more bigger platforms. It will be important to provide flexibility through a highly innovative ecosystem that enables the emergence of more complex sets of services and platforms (including free and commercial services).

Platforms and clouds help to reduce the complexity on the user side (in this case mostly companies and organisations that build products or platforms on top of the enabling services) and support evolution (competition and cross-fertilisation) on the service provider side. Our goal is to fully adopt, in the Language Technology community, the successful approach of hybrid research or DevOps, i.e., a tight and integrated loop of research, development and operations that allows for early testing and short development cycles.

Users of the technologies will be able to receive customised integrated services without having to install, combine, support and maintain any software. They will have access to specialised solutions even if they do not use these regularly. *Language technology providers* will have ample opportunity to offer stand-alone or integrated services through component technologies or cloud-based APIs. *Researchers* will have a virtual laboratory for testing, combining, and benchmarking their technologies and for exposing them in realistic trials to real tasks and users. Through the involvement of *users*, valuable data will be collected within these inherently European platforms (vs. platforms that physically reside on other continents) that can directly feed back into improved services. *Providers of services* that can be enabled or enhanced by text and speech processing will utilise the platform for testing the needed LT functionalities and for integrating them into their own solutions. *Corporate users* will enjoy the benefits of language technology early and at no (or reasonable) cost through a large variety of generic and specialised services offered through a small number of sources.

In order to allow for the broad range of potential solutions, the emerging platforms will have to host (and share) all relevant simple services, including components, tools and data resources, as well as various layers or components of higher services that incorporate simpler ones. Resource exchange

infrastructures such as, for example, META-SHARE (including the Linguistic Linked Open Data approach) will play an important role in the design of the platform.

The initial design and creation of these services and platforms has to be supported by public funding. Because of the demanding requirements regarding performance, reliability, user support, scalability, and persistence together with data protection and compliance with privacy regulation, the systems need to be established by one or more consortia with strong commercial partners and also be operated by these consortia or commercial contractors.

2.2. Application Examples

In the following we illustrate the functionality of the foreseen applications with a few application examples: services are needed that provide flexible multilingual technologies – even including *human* translation. These services need to be designed from the outset with special emphasis on high-quality output, trust, data security, reliability, privacy, data protection and confidentiality. We also need a bridge to the world of knowledge, data and meaning through corresponding services. This bridge needs to provide seamless and ubiquitous access to multilingual knowledge bases that integrate information about products, companies, places, terms, words, and a plethora of other concepts that are important for all monolingual, crosslingual and multilingual language technology components and data value chains. The design and implementation of such a knowledge service is a challenge but it can become a reality through the combination of repositories such as Wikipedia, Wikidata, BabelNet, DBPedia, Linked Open Data sets and other resources and data sets. Important for industry and e-government will be sector-specific multilingual knowledge systems, which are an essential prerequisite for serving a global customer base. Another candidate for one of the sets of generic services is concerned with sophisticated methods for text analytics and production, such as, for example, report generation, text classification, sentiment analysis and opinion mining. As a third example, we foresee conversational technologies and natural language interaction services for dialogue systems and interactive voice interfaces that include the analysis and synthesis of spoken language. All these services would need to be linked up with one another. The spoken language services, for example, would need to contain bridges to the translation services. All services would need to have a 24/7 availability and provide web-scale performance and connectivity to carry out their main purpose: support and enable the application solutions and to support research and innovation by testing and showcasing results as well as providing an environment for hybrid research and devops, i.e., integrating operational services and research.

The MLV Programme will unlock the Multilingual Digital Single Market through a set of services, platforms and applications that support all businesses and citizens. It will also provide the European language technology community and several different industries with the ability to compete with other markets and achieve multiple benefits for the European economy and future growth, as well as for society and the citizens.

To achieve this ambitious plan, all stakeholders need to collaborate and cooperate closely and in a tightly coordinated and efficient way. To demonstrate that the whole Multilingual Europe community firmly stands together, this document is presented by the Cracking the Language Barrier federation, which – at the time of writing (July 2016) – consists of 10 organisations and more than 20 projects already working together on the technological foundations of a Multilingual Europe, some of them as early as 2010 (META-NET).

2.3. Timeframe and Costs

The MLV Programme foresees three highly focused years of work which correspond to three phases (2018–2020). We recommend to start with the implementation of the MLV Programme at the beginning of 2018. The time until then is needed to come to a consensus with all stakeholders involved regarding the main aspects and priorities of the programme.

The key principle of the MLV Programme is a solid, robust and operational **Multilingual Services** layer, on top of which **Multilingual Applications** can be realised as innovative solutions. We can start with the further design and implementation of the Multilingual Services quite rapidly due to significant investments of the EU into language technologies topics in the last ca. 15 years. Even with a partially

established, incomplete set of Multilingual Services, we can start building Multilingual Applications already in 2018. At the same time, we can start bringing in new results from **Research** into the set of Multilingual Services to provide revised, improved or completely new, additional or alternative services. The three-layer-approach does *not* mean that we have to invest in research first to harvest the results later – quite the contrary, i.e., activities on all three layers can be initiated at the same time. In that regard, it is important to note that we anticipate projects not to focus upon one single layer but to address two layers at the same time (**Research and Multilingual Services** or **Multilingual Services and Multilingual Applications**); several larger actions may address all three layers.

This decentralised and decoupled approach with an evolutionary growing set of multiple services and platforms that can be developed independently from the actual applications is the appropriate enabler for an agile multilingual value ecosystem, which is continuously driven by highly agile and innovative research projects. The significant overlap between the European multilingual value ecosystem and the European data value ecosystem is intended.

As soon as there is agreement between all stakeholders involved in our proposal, including, crucially, the funding agencies, the next step is to further specify the exact timing and roadmap of the three years and phases (2018, 2019, 2020). A prioritisation of the Multilingual Services and Multilingual Applications, most importantly needed especially in the important first two years is also among the next steps, to be included in the final version of this document. The same holds for the identification of the most important and most urgently needed services as well as applications. Our current proposal does not include any specific priorities yet because these need to be discussed and identified in a multi-stakeholder dialogue.

An important task of the first three years will be the conceptualisation of business models, especially around the Multilingual Services (B2B) but also around the Multilingual Applications (B2C, B2B). We expect the creation of a multitude of multilingual services which need to be seed-funded and then transformed from projects into their own entities or products to guarantee independent and sustainable operation after 2020. We expect the increased adoption of new and innovative business models for these services that will involve the creation and acceleration of many startups and spin-off companies in Europe.

In the following we list the main aspects of the three phases of the MLV Programme including a pre-programme phase. We expect the MLV Programme to result in a set of technology solutions, applications and services, effectively realising the Multilingual Digital Single Market.

- **Pre-MLV (2016/2017):** stakeholder discussions and consensus building; finalisation of the strategy and roadmap; final selection and prioritisation of topics; set up of several smaller prototype and proof-of-concept projects as well as support actions for planning
- **MLV Phase 1 (2018):** establish first set of Multilingual Services; conceptualise Multilingual Applications; foresee integration of services into existing applications; continue activities in the priority research themes
- **MLV Phase 2 (2019):** further extension of Multilingual Services (coverage, quality, precision) and intensified standardisation activities; deployment of first applications; work on business models; continue activities in the priority research themes
- **MLV Phase 3 (2020):** further extension of Multilingual Services (including standardisation); deployment of Multilingual Applications; transformation of projects into independent and sustainable entities; continue activities in the priority research themes
- **Post-MLV (2021+):** Scaling up and extending the Multilingual Applications and Multilingual Services; expanding language and domain coverage; going beyond Europe, penetrating other markets; exploration of novel research strands etc.

The total estimated costs for a basic implementation of the MLV Programme with a set of mission-critical services and applications is in the range of 175-200 Million Euros for the first three years and phases (2018, 2019, 2020), including industry contribution (ca. 20%).

After these three years, we expect a reduction of funding for the Multilingual Services and the Multilingual Applications layer – the goal is for the Multilingual Services to become independent, self-

sustained and profitable; the Multilingual Applications are meant to include a significant commercial component right from the start. At the same time, funding for Research needs to be kept at least at the same level, ideally even increased, in order to secure a leading position for our continent in this highly important field.

3. Application Roadmap for the Multilingual Digital Single Market

In this chapter we provide the first version of an Application Roadmap for the Multilingual Digital Single Market. The list of multilingual applications and services is not meant to be complete but constitutes an important set of needed components. It reflects the current state of discussion within the Language Technology community.

Despite the planned technological independence (see the previous Chapter), certain technologies can be foreseen for either the Applications or the Services layer – maybe even for both. Most of the technologies listed below relate to the Multilingual Applications layer, some relate to the Services layer. These need to be further specified on the way to the final version of this strategic agenda document.

3.1. Application Area: E-Commerce

3.1.1. The Customer-facing Side

- **2018:** Technologies and workflows for multilingual Content Management Systems
- **2019:** Multilingual and crosslingual search engines, product aggregation, product/service comparison and IR with a focus on European languages
- **2020**: Automatic translation of online shops and other websites (including semi-automatic localisation and internationalisation) to enable SMEs to offer their services in more languages

3.1.2. The Back-Office-facing Side

- **2018:** Creating domain-specific, multilingual vocabularies, taxonomies, ontologies, product catalogues etc.
- **2019:** Seamless integration of content, data, and knowledge across multiple modalities and channels (mobile, web, voice etc.), incorporating open and closed datasets in a way that is respectful of intellectual property (IP), privacy, and licenses.
- **2020:** Unified customer experience and cross-cultural CRM through multilingual and crosslingual technologies, pushing contextual, up-to-date and relevant additional information about products or services to users, bringing together content, customer care, CRM, discussion fora, helpdesks (including sentiment analysis) etc. in a unified digital (eco)system across languages

3.1.3. Crosslingual Semantic Product Descriptions and Catalogues

- **2018:** Shared ontologies and terminologies for key administrative, financial and legal sub-domains, to foster the creation of the Digital Single Market and the goals of CEF
- **2019:** Automatise the generation of (inherently domain-specific) machine-readable product descriptions, to foster personalised cross-border and cross-lingual product discovery, aggregration and comparison

3.1.4. Online Dispute Resolution (ODR)

- **2020:** Support the EC's ODR platform by identifying similar cases and semantic links between similar cases; enable merchants/service providers and consumers to settle their disputes outside courts, across borders, in situations where they do not have a common language. Needed are MT solutions, a glossary database, spell checkers, translation of free text fields and different types of languages (formal, informal etc.).

3.2. Application Area: Content, Media, Verticals

3.2.1. Analytics, Curation, Generation, Authoring Support

- **2018:** Ensure that data is ready for sustainable commercial exploitation, including quality information of automatically translated content and data provenance
- **2018:** Multilingual and multimodal text, data and media analytics (including NER, temporal analysis, linking etc.).
- **2019:** Solutions to support the multilingual, automatic or semi-automatic generation of articles and reports based on Big Data, linked data and other sources of structured or semi-structured information
- **2019:** Analysing user feedback (identifying semantically similar communications) for more efficient cross-lingual communication in customer support.
- **2019:** Intelligent cross-lingual authoring and enrichment of content, including linking to other content objects or semantic concepts, and making it understandable across language barriers and for machines and humans alike; including authoring support that can flag potential errors, suggest corrections, and use authoring memories proactively to suggest completions of started sentences or paragraphs; advanced technologies can check for appropriate style according to genre or text type and help improve comprehensibility.
- **2020:** Enable and support automatic repurposing of media content across languages (including discoverability, metadata, licensing)
- **2020:** Services for novel content curation approaches like semantic storytelling to foster new business models in the content and media industry, including advanced algorithms that can adapt perspective, tone, and humour to tailor a story to its audience.
- **2020:** Solutions for enabling the semantic interoperability of data sources to help extract and combine content from multiple data sources and across all communication channels (telecommunication, meetings, email, chat etc.)
- **2020:** Structured data analysis to identify facts and trends, combining them with contextual information to form and string together sentences, enabling the (semi-) automatic generation of multilingual articles, reports or websites.

3.2.2. Multimodal Language Interfaces

- **2018:** Generic written- and spoken-language interfaces for the IoT (smart phones, cars, consumer products, household appliances, chat-bots); includes gesture- and text-based multimodal interfaces as well as smart multilingual assistants embedded in wearables.
- **2019:** Automated interpretation solutions for spoken communication including automatically transcribing, translating and eventually summarising the content of meetings (including incrementally drafted summaries for displaying the state of the discussion, intermediate results and open issues, and to generate meeting minutes.)
- **2020:** Enhanced virtual meetings through multilingual or crosslingual virtual meeting rooms that provide services like seamless spoken translation or automatic note taking.

3.2.3. Verticals: Health, Legal, Government, Mobility, Energy

Especially with regard to the multilingual applications that relate to these crucial societal vertical domains there is a very close connection to the topics of the "Translation, Language, Knowledge, Data" application area, mentioned below.

- **2018:** E-government and legal aspects: terminologies, linked data sets, and ontologies that harmonise the concepts used in different countries and jurisdictions, as a basis to reach interoperability and to develop a new generation of (public) services that is implemented across countries with multilingual technologies built in (including provenance and licensing information).
- **2018:** E-health: create a single market for health services where practitioners, patients, and administrators can communicate seamlessly across language barriers; tools and methodologies

are needed for HQ translation, codes need to be extended to all EU languages; automatic translation reliability is needed for e-health/medical concepts and terms as defined and modelled by terminologies in a given EU language but also to be understood in the medical domain and/or a given health system context.

- **2019:** E-government: solutions to further improve the pan-European cross-border exchange of electronic documents, cross-border communication including legal aspects, specialised free translation services – towards a borderless e-government space in Europe (in collaboration with and supporting CEF).
- **2020:** E-procurement platform in which multilingual LT supports the translation of user interfaces, documents, and larger narratives (currently performed manually). LTs are needed for concept identification and extraction, matching offer and demand to identify business opportunities and to produce accurate summaries for decision-making with regard to Calls for Tenders.

3.3. Application Area: Translation, Language, Knowledge, Data

3.3.1. Translation Centre

- **2018:** Open, comprehensive, integrated platform for machine and human translation; generic and specialised federated services for instantaneous reliable spoken and written translation among all European and major non-European languages
- **2018:** Tools, systems, workflows and infrastructures for bridging human translation and machine translation; different service layers (public, internal); can include a free 24/7 public service of basic automatic services (text translation, term and word translation); professional services for a fee; integration of CEF.AT.
- **2018:** Integrating translation technologies (also supporting niche languages and micro-domains) into systems will enable companies to build more efficient, crosslingual CRM systems; allows micro-SMEs in ecommerce to exploit multilingual value chains.
- **2019:** High-quality and high-speed machine translation for many languages, across multiple subject fields and text types, both spoken and written; including customisable MT services, written and spoken language as well as solutions for specialised micro-domains (including languages, cultures, measurement systems, safety regulations, work habits etc.).
- **2020:** Automatic localisation and translation of ecommerce text types including documentation, instructions, manuals, insurance, privacy protection, validation forms, after-sales information.

3.3.2. Text, Data and Speech Analytics

- **2018:** Multilingual text and speech analytics for Public and Business Intelligence
- **2019:** Technologies to monitor, analyse, summarise, translate, structure, document, and visualise social media dynamics and enable multilingual and cross-lingual Public and Business Intelligence market research.
- **2019:** Analysis of large volumes of data to generate summaries, detect trends, answer questions, and search concepts etc. for significantly improved Public and Business Intelligence (including question answering and semantic search); including general and domain-specific approaches at topic extraction, document classification, entity recognition, relation extraction, event extraction.
- **2020:** Voice of the Customer and Voice of the Citizen: multilingual market/society research including extracting and interpreting the multilingual voice of the customer/citizen with a high level of accuracy, across languages and modalities (including web-scale sentiment analysis, and opinion mining at deeper levels beyond mere polarity, including intention recognition); including summarising multilingual data streams in real time, dealing with high-volume, high-velocity data, often of unknown veracity, opinion mining, multilingual report generation, trend analysis.

- **2020:** Analysis of the formation of collective opinions and attitudes including the detection and analysis of trends and emotions; the analysis of sentiment at deeper levels is a crucial component of social intelligence.

3.3.3. Language, Knowledge and Data

- **2018:** Large-scale dynamic multilingual knowledge graphs including LOD and domain-specific vocabularies, taxonomies, ontologies, data sets
- **2019:** Achieve scalable creation, discovery and exploitation of multilingual public sector information (PSI) data sets for re-use of PSI across languages and countries, implementing the EC directive of public information reuse
- **2019:** Self-contained, adaptable, flexible, open platform for generic, pluggable, configurable data, language and knowledge services that can grow into a larger ecosystem, also including Big Data platforms and approaches; focus upon engineering and sustainability, not on the actual LT research; infrastructure should be able to support a wide range of services, e.g., basic low-level technologies such as POS tagging and high-level (combined) ones such as MT including special terminology and human post-editing, generation of spoken usage instructions, or email classification by sentiment and enrichment with background information.
- **2020:** Generate rich, linked knowledge resources for multimodal and multilingual repurposing of heterogeneous content for different challenges, natural languages, and audiences (including linking resources, visual story generation from multimodal data, semantic user profiles). Linked Data to create a unified information space by bringing together heterogeneous data including product data, customer data, and social data.

3.3.4. Language Resources

- **2018:** Ease the re-use of linguistic resources in all parts of the data value chain across languages and sectors.
- **2018:** Basic monolingual technologies and resources for Multilingual Europe
- **2018:** Next generation LR/LT infrastructure to cope with current requirements
- **2019:** Allow for the joint exploitation of public and private data sources
- **2019:** Automatize the creation of data needed for multilingual and cross-lingual semantic annotation scenarios in a scalable and sustainable manner
- **2020:** Multilingual technologies for Multilingual Europe, especially for under-resourced languages

4. Background Information: Selected Applications and Services for the Multilingual Digital Single Market

In this chapter we provide background information on selected multilingual applications and services for the Digital Single Market, foreseen to be developed in the MLV Programme. The end-user applications are meant to be developed, first and foremost, in collaborative European projects by commercial solution providers and European R&D&I, based on the MLV services and platforms. The primary customers of these solutions are, again, companies (primarily SMEs) but also public administrations, European institutions, the European citizens and other stakeholders interested in the multilingual Digital Single Market.

4.1. Multilingual Application: E-Commerce, CRM and After-Sales

- Applications for crosslingual ecommerce: automatic translation of online shops and other websites (including semi-automatic localisation and internationalisation) to enable SMEs to offer their services in more languages, penetrating the whole multilingual DSM
- Provides contextual, up-to-date and relevant additional information about products or services to users, bringing together content, customer care, customer relationship (CRM), discussion fora, helpdesks etc. in a unified digital (eco)system *across languages*

Nowadays consumers expect to quickly and easily get what they need from a business – anytime, anywhere. This includes access to products and services, but also to information and easy-to-use self-services. Industries interact with their customers on a daily basis. They have to recognise customer needs and intentions in real-time and guarantee the consistency of provided information across channels, audiences *and languages*. Automation helps bring together content, product and customer relationship management in one ecosystem. The goal is a seamless network of content, data and knowledge that spans multiple modalities and channels (mobile, web, interactive voice response etc.) and incorporates open and closed datasets in a way that is respectful to intellectual property (IP), data privacy and licenses. Realising agility at the content level will enable the quick integration of new (external) data resources and will allow marketing experts to dynamically react to changing customer and market needs. Linked Data technologies can help create a unified information space by bringing together data from different sources, including product data, customer data, and social data. The generation of rich linked knowledge resources enables multimodal and multilingual repurposing of heterogeneous content for different challenges, natural languages and audiences. Linking resources can enable the visual story generation from multiple sources including text, video and other modalities, or the creation of semantic user profiles based on linked information about objects, individuals, groups, intentions, contexts, and cultures. In cross-cultural CRM, the integration of translation technologies will enable companies to efficiently engage with their customers across languages. This will not only allow micro-SMEs in ecommerce to exploit multilingual value chains, making them competitive in market niches, but also help create an extraordinary, contextualised digital experience to all users.

4.2. Multilingual Application: Business Intelligence using Big Data

- Analyse large volumes of data, generate clear summaries, detect trends, answer questions, and search concepts (instead of words) etc.

Language makes (very) big data in the web, intranets, and various databases, user fora, among others. However, although much of the most relevant information is contained in texts, text analytics applications today only account for less than 1% of the more than US$ 10 billion business intelligence and analytics market. Because of their limited capabilities in interpreting texts, mainly business news, reports and press releases, their findings are still neither comprehensive nor reliable enough. While the main tasks required in text analytics are rather conventional (topic and event extraction, document classification, entity extraction, relation extraction), there is a clear need for analytics solutions that are tuned to the needs of particular domains and are able to generate and incorporate semantic domain-specific knowledge in the form of taxonomies or terminologies to support domain customisation. Summarisation technologies as well as multilingual report generation will help keep up with linguistic big data to be processed. In the other direction, semantic search, question-answering and trend detection services will allow business analysts, decision makers, venture capitalists and other experts to access complex information in a targeted way. Adaptive techniques are needed to ensure that the responses to the user are relevant: choosing the correct information to provide for that context, and automatically phrasing it in a relevant way that supports their need for further explanation, illustration or clarification.

One of the technical priority themes identified in the Big Data Value Strategic Research and Innovation Agenda is Deep Analytics. Several of the major expected advanced analytics innovations, such as semantic analysis and multimedia (unstructured) data mining, strongly relate to natural language data. Language technology will be the foundation for organising, analysing, and extracting big data in a truly useful way. To this end, we suggest collaborating closely with the BDV cPPP to provide this set of technologies.

4.3. Multilingual Application: Online Dispute Resolution

- Multilingual support for the EC online service where merchants/service providers and consumers can settle their disputes outside courts, across borders, in situations where they do not have a common language.

The EC is devising and deploying an interactive and free-of-charge website for Online Dispute Resolution (ODR). ODR aims at resolving contractual disputes from European consumers (B2C) or traders (B2B), which arise from cross-border and domestic online sales or service contracts. Competing for alternative dispute resolution (ADR) models requires not only managing the translation of messages, conversations/mediations flowing among parties: evaluating, sending and receiving information, but also translating documents needed for finding a resolution to the dispute and other needed functionalities of the platform (guidance, easy to fill forms, etc.). We suggest, thus, that ODR uses machine translation to provide a multilingual platform. The EC has a legal obligation (Regulation on consumer ODR and Directive on consumer ADR) to implement this platform in all official languages of the institutions of the EU. The ODR platform presents a unique opportunity. Main multilingual challenges of the platform comprise managing 500 language pairs, a glossary database, spell check, translation of free text fields and different types of languages (formal and everyday languages). The ODR platform not only poses significant challenges for LT; a high-quality multilingual tool could underpin the uptake and credibility of LT among customers and users. The ODR system aims at boosting online purchases from consumers and traders (especially at cross-border level); visibility of LT could exponentially increase as the ODR platform will be accessible to millions of consumers and thousands of traders using ecommerce in Europe. A close collaboration with the CEF Automated Translation activity is foreseen.

4.4. Multilingual Application: Voice of the Customer and Voice of the Citizen – Social Intelligence on Big Data

- *Voice of the customer:* Multilingual market research, i.e., extracting and interpreting the multilingual opinions with high accuracy, across languages and modalities, and analysing sentiment as well as opinion at deeper levels beyond mere polarity (including intention)
- *Voice of the citizen:* the same set of technologies but applied to social, sociological, ecological and other related, non-commercial aspects; the goal is to enrich democracy by new mechanisms for improved collective solutions and decisions (also see E-Participation).
- Technically: large-scale, web-scale sentiment analysis, opinion mining, multilingual report generation, trend analysis

The recognition of user needs, intentions and opinions towards products and services is crucial for the success of today's companies. Recognising customer needs and opinions involves the extraction and interpretation of customer interactions with a high level of accuracy and across languages and modalities. The main source of information is user-generated content from social media. Customers and potential customers share their thoughts using blogs (e.g., Twitter), post comments in online forums, or send feedback via email. All these text and voice messages are a valuable source for trending sentiments and opinions about products and services. We foresee technologies for the (targeted) analysis of large volumes of such comments and communications created by citizens, customers, patients, employees, consumers and other stakeholder communities. Summarising multiple multilingual data streams in real time involves dealing with high-volume, high-velocity data, often of unknown veracity. As the formation of collective opinions and attitudes is highly dynamic, new developments need to be detected and trends analysed. As emotions play an important part in individual actions such as voting, buying, supporting, donating and in collective opinion formation, the analysis of sentiment at deeper levels is a crucial component of social intelligence. Text analytics will play a role in areas such as analysing the voice and actions of the customer in the context of CRM; brand, product and reputation management; technology monitoring and competitive intelligence. Automatic curation, summarisation and translation technologies will help monitor, analyse, summarise, structure, document, and visualise social media dynamics and enable multilingual and cross-lingual market research. Technologies such as sentiment analysis, opinion mining, and intention recognition will extract and interpret the voice of the customer and also that of the citizen with a high level of accuracy and across natural languages and modalities, while providing insights how culture and behaviour affect any conclusion. This technology solution also includes the application of the abovementioned methods to spoken language data, collected, for example, in (automated) call centres.

Following the Fukushima incident in 2011 there have been discussions about the dangers of nuclear energy in all European countries. These debates were held in the respective language communities only, there has never been a public European debate about the topic because it is, technically, not yet possible to organize such a debate online. The "Voice of the Citizen" application is intended to help pave the way to full e-participation by providing technologies for multilingual social media mining. The idea is to analyse social networks and user generated content in all European languages in order to gather concrete numbers and statistics about what Europeans in specific countries or regions think about urgent or important topics such as e-mobility, nuclear energy, climate change etc. Such information can be used to inform European decision support, to increase social reach and also to improve cross-cultural understanding. The goal is to create a "citizen experience" – as a complement to the unified "customer experience" or "user experience" for commercial products, services or offerings.

4.5. Multilingual Application: Content Curation and Content Production

- Support the intelligent authoring, enrichment, linking and processing of content, making it readable and understandable across language barriers and for machines and humans alike
- Support the multilingual, automatic or semi-automatic generation of reports and articles based on Big Data, Linked Data and other data sets

Collecting, organizing, structuring and displaying information relevant to a particular topic or area of interest is a major task in many areas, including journalism, marketing and decision-making. Accelerating the process of discovering relevant content is especially crucial for those whose work involves the processing of large amounts of information in a short time. Technologies for digital content curation reduce the overall flow of information and make them more targeted to the end user's interests. Machine translation technology can help handle multilingualism of data sources and facilitate access to multilingual data assets. Semantic technologies are crucial for enabling the semantic interoperability of data sources and help extract and combine content from multiple data sources and across all communication channels (telecommunication, meetings, email, chat etc.). Technology can also be of help in various content production tasks. Standardised communication, e.g., email communication in customer support, can be automated by analysing user feedback and identifying relevant, semantically similar previous communications. Robot journalism can comb structured data for facts and trends and combine them with contextual information to form and string together sentences, enabling the generation of multilingual articles, reports or product websites. Advanced algorithms can adapt perspective, tone, and humour to tailor a story to its audience. In human text generation, authoring support software can flag potential errors, suggest corrections, and use authoring memories proactively to suggest completions of started sentences or whole paragraphs. Advanced technologies can check for appropriate style according to genre and purpose and help improve comprehensibility.

4.6. Multilingual Application: Written- and Spoken-Language Interfaces

- Robust written- and spoken-language interfaces and dialogue systems for connected devices (including chat bots for web applications)
- Bridge to the Internet of Things (IoT), Web of Things (WoT) and Industrie 4.0 (Advanced Manufacturing) area for which voice interfaces will become the norm in the near future

The number of connected devices is continuously growing (Internet of Things, Web of Things). Depending on their function and complexity, the nature of desired or needed communication can vary widely. Some objects will come with interesting textual information (manuals, consumer information), others will provide information on their state and will have their own individual digital memory. Objects that can perform actions, such as vehicles and appliances, will accept and carry out (multilingual) voice, gesture or eye-tracking commands. Wearable sensors can provide signals about a person's mood or emotional state, offering new affect-focused multimodal interaction with devices. In addition, robots are now evolving into collaborative, social machines that will eventually provide useful services to humans in numerous work, medical, educational and household contexts. Dialogue systems and multimodal conversational interfaces that support natural language commands have the potential

to adapt automatically to the user and to the environment. For instance, interfaces adapted to elderly people will take into account cognitive, auditory, visual, and articulatory ageing; interfaces will adapt to what a user is doing (working in a noisy, hands-free environment, e.g., when rushing for a train); systems for devices where "traditional" interfaces (keyboard, mouse, trackpad, touchscreen, etc.) are not usable (e.g., small wearable devices) or not appropriate ("companion" systems); smart mobile agents which are capable of deeper natural language and multimodal interaction, possibly focused on specific domains, and capable of rich question answering. Many vertical market sectors with domain-specific assistants exist: shopping, travel, social service planning, learning and tutoring.

4.7. Multilingual Application: Translation Centre

- Customisable machine translation services including written and spoken language as well as solutions for specialised micro-domains
- Broad set of target users: businesses, governments, administrations, customers, citizens
- Broad set of use cases: from desktop to mobile to tablet to voice to automatic (via API)

Translation services are moving to cloud-based solutions – generic and specialised federated services for instantaneous reliable spoken and written translation among all European and major non-European languages. Clouds make it possible to offer different service layers such as a public and an internal service layer for providers with different offerings. This can include a free 24/7 public service of basic automatic services (text translation, term and word translation), professional services available for a fee (including high-quality professional services by human translators, terminology, dictionaries, checking, TMs) and free human translation or post-editing services for special purposes provided by NGO-initiatives. The Translation Centre foresees one common, easy-to-use access point for citizens, professionals, businesses, and public organisations providing ubiquitous and instant access to information and communication in any language. Behind this access point will be a network of generic and special-purpose services combining automatic translation or interpretation, language checking, post-editing, as well as human creativity and quality assurance, where needed, for achieving the demanded quality. For high-volume base-line quality the service will be free for use but it will offer extensive business opportunities for a wide range of service and technology providers.

When they travel across borders, products and services are typically tailored to foreign communities and accompanied by documentation covering instructions, insurance, privacy protection, validation forms, after-sales information and more. All this content needs to be adapted to the languages, cultures, measurement systems, safety regulations and work habits of new customers and end users. Systems need to be engineered to automatically control this process, cut lead times, radically reduce transaction costs, and improve information quality. A technological solution will be critical for the multilingual Digital Single Market.

High-Quality Machine Translation (HQMT) in the cloud will ensure and extend the value of the digital information space in which everyone can contribute in her own language and be understood by members of other language communities. It will assure that diversity will no longer be a challenge, but a welcome enrichment for Europe both socially and economically, especially with regard to the multilingual Digital Single Market. As a tool for engaging online with the richness of cultures across Europe, HQMT can act as a doorway to, rather than a substitute for, acquiring the multilingual skills needed to travel and immerse in other cultures more fully. Some of the showcase applications include multilingual content production (media, web, technical, legal documents), cross-lingual communication, document translation and search, real-time subtitling and translating speech from live events, mobile interactive interpretation for business, social services, and security, translation workspaces for online services.

High-quality services require a combination of Human Translation (HT), Computer-Assisted Translation (CAT) and full Machine Translation (MT). Core requirements are trust in the reliability and accuracy of translation and the security of the translation channel. The platform should include translation companies and experts in a variety of domains (e.g., bio- medical, financial, legal, scientific), tasks and genres (e.g., technical documentation, business reports, fiction, etc.); extended services like multilingual text authoring, multimedia translation, and quality assurance by experts;

mechanisms for customer care and trust building; certified security systems; quality upscale models (instant quality upgrades). In this platform a close collaboration and bridge to the MT@EC system (CEF AT) is foreseen. In addition to the EC MT services, similar services can be provided by private companies, research centres, universities or NGOs such as Translators without Borders or the Rosetta Foundation. The platform could be operated by an industrial interest group (EEIG) in close collaboration with MT@EC. Necessary ingredients are a powerful and stable service and service brokerage platform with an API to automatic or quasi-real time human services provided by a set of initial LSPs and MT systems. It could be hosted by trusted service centres, i.e., certified service providers fulfilling highest standards for privacy, data protection, confidentiality and security of data and translations.

4.8. Multilingual Application: E-Government

- Improve – in close collaboration with CEF – the pan-European cross-border exchange of electronic documents, cross-border communication including legal aspects, specialised free translation services – towards a borderless e-government space in Europe.

The creation of a multilingual Digital Single Market should also include vastly improved cross-border public and government services that interoperate and counter market fragmentation, in particular in the areas of e-government, e-health and e-procurement. This set of solutions foresees, among others, e-procurement platforms in which multilingual language technologies can support the translation of user interfaces, documents and large narratives that are currently performed manually. Language technologies are needed for concept identification and extraction, matching offer and demand to identify business opportunities and to produce accurate summaries for decision making in tendering processes.

Many of the technologies to be developed through the MLV Programme can also be used effectively in e-government scenarios, especially sophisticated high-quality machine translation methods, or text analytics technologies. Specific to e-government are the development of terminologies, linked data sets, and ontologies that harmonise the concepts used in different countries and jurisdictions, as a basis to reach interoperability and to develop a new generation of services that is implemented across countries with multilingual capabilities built in. We suggest to design and to deploy an ecosystem of data that is partially open and partially closed but is extended with appropriate provenance and licensing information as well as mechanisms for representing and dealing with trust and confidence, so that the public as well as private companies can exploit the data for their purposes and within their applications. We also need technologies to generate reports and reviews automatically. Automatic processes can take raw data to transform the numbers and words into succinct reports for later use by specialists. This will save time and money and rapidly inform all stakeholders for further discussion.

4.9. Multilingual Application: E-Health

- Cross-border healthcare scenarios open new ways for creating a single market where practitioners, patients and administrators can communicate across language barriers.

When considering cross-border health care, it was shown that challenges go beyond the technical level and include different interactions with health professionals and patients. Interoperable e-health systems need not only different interfaces to manage data, text or speech, but also to cover different challenges in different levels of the data value chain. Tools and methodologies are needed for high quality translation, codes need to be extended to all EU languages. Automatic translation reliability is needed not only for e-health/medical concepts and terms as defined and modelled by terminologies in a given EU language but also to be understood in the medical domain and/or a given health system context.

4.10. Multilingual Application: E-Learning

- Life-long learning and multilingual online training courses will help self-studying, cross-border migration, training for staff members of pan-European companies etc.

The software market for computer-assisted language learning (CALL) is growing fast. While current products can help traditional language instruction, they are still limited in functionality because the software cannot reliably analyse and critique the language produced by the learner. This is true for written language and even more so for spoken utterances. Software producers are trying to circumvent the problem by closely restricting the expected responses of the user, something that helps for many exercises, but still rules out the ideal interactive CALL application: an automatic dialogue partner ready around the clock for error-free conversation on many topics. Such software would analyse and critique the learner's errors and adapt its dialogue to the learner's problems and progress. Current language technologies cannot provide such functionality yet. Its lack of flexibility is the reason why research on CALL applications has not yet come into full bloom. As research on language analysis, understanding and dialogue systems progresses, we can predict a boom in the promising and commercially attractive CALL area. However, use cases for language technologies in this area are much broader. Machine translation can help in accessing massive open online courses (MOOCs), virtual assistants can help in tests and learning, language technologies are also essential for multilingual gamification.

4.11. Multilingual Service: Knowledge and Data Repositories

This platform and set of repositories will bring in services for processing and storing knowledge gained by and used for understanding, translation, curation and generation. It will include knowledge graphs, linked data sets and ontologies, as well as services for building, using and maintaining them. The goal is not to model arbitrary world knowledge but rather to realise selected forms of inference needed for utilising and extending knowledge, for understanding and for successful communication (including better decision support, pro-active planning and autonomous adaptation). The W3C standards for creating, managing, interlinking and searching the open data of the web have matured to the level that they can fully support open, massively multilingual language resources that integrate semantic knowledge, lexical knowledge, annotations, corpora, online content and data sets of all types. Several open source tools already exist and there is a rapid migration of language resources to this technological platform.

The goal is to base the platform and repository fully upon the Linked Data paradigm to ensure that data and services form a linked ecosystem rather than a set of fragmented and non-interoperable datasets. Standardised vocabularies ensure convergence. Technologies conformant to web standards (RDF, OWL) offer powerful APIs such as SPARQL for search and RESTful services to publish, update and manipulate linked data on the web. De-centralisation is key in that the implementation of the architecture is web-based and does not rely on any central node or service nor on particular providers of a cloud. In particular, this should prevent any vendor lock-in and dependencies on particular agents.

This resulting Linguistic Linked Data (LLD) platform is structured as follows: Multilingual data, in all forms, modality and media types are the foundation (with mappings to XML vocabularies, JSON and CSV). Metadata provide information about datasets (author, language, structure), etc. including license information and description of copyright information and other rights-related restrictions (e.g., privacy and data protection of personal data). Of utmost importance is the provenance of the data, i.e., its origin and processing history. Additionally, the platform needs to integrate functionalities for the consistent description, publication and inter-linking of resources (lexica, corpora, terminologies etc.), using, ideally, common vocabularies. Furthermore, specific single and also composable services need to be specified and implemented in an interoperable way in order to bridge between the knowledge platform and between language analytics as well as other processing services in terms of producing or consuming data and knowledge from the platform on a web-scale level.

Many elements of the platform are already in place, based on open source tools, specifications and guidelines from the LOD2 stack and the W3C Data Activity. Guidelines and tools specific to linguistic linked data are being actively promoted by several W3C, LT and Knowledge communities. Massively

multilingual examples of aggregation and discovery solutions using LLD are publically available and already have a major impact on the NLP and language resource communities. Babelnet[20] aggregates lexical-conceptual information from Wikipedia, Wikidata, different Wordnets and Wiktionary into a single service supported by the annotation service Babelfy. It covers 271 languages, 117 million lexical senses, over 6 million concepts, 7 million named entities, 10 million images, all interlinked by 354 million lexico-semantic relations using nearly 2 billion RDF triples. The Linguistic Linked Data Cloud[21] of interlinked resources is now an important and growing part of the overall LOD cloud. The language resource community is well on the way to wholesale adoption of linked data as its primary data exchange mechanism.

4.12. Multilingual Service: Language Processing, Analysis and Production – Language Resources

An essential, important prerequisite for all service and platform activities is pooling and sharing data sets, language resources and technologies. In this regard, one of our key goals is to set up a shared initiative together with, ideally, all EU Member States and all interested associated countries, in order to collaborate closely with all national and regional research centres and universities, thereby making use of their respective expertise vis-à-vis their own national or regional languages in terms of language technologies and, maybe even more important, computational modelling and computational linguistics methods for automatic language processing and production. A similar approach is currently followed by CEF AT with dedicated service contracts to identify and collect data sets in the CEF-participating countries.

The three groups of Multilingual Services share a large and heterogeneous group of core technologies for language analysis and production that provide development support through basic modules and datasets. To this group belong tools and technologies such as, among others, tokenisers, part-of-speech taggers, syntactic parsers, tools for building language models, IR tools, machine learning toolkits, speech recognition and speech synthesis engines, and integrated architectures such as, among others, GATE, UIMA and FREME.

Many of these tools depend on specific datasets (i.e., language resources), for example, very large collections of linguistically annotated documents (monolingual or multilingual, aligned corpora), treebanks, grammars, lexicons, thesauri, terminologies, dictionaries, ontologies and language models. Tools and resources can be rather general or highly task- or domain-specific, tools can be language-independent, datasets are, by definition, language-specific.

A key goal of the MLV Programme is to collect, develop and make available core technologies and resources through a shared infrastructure so that the research and technology development carried out in all themes can make use of them. Over time, this approach will improve the core technologies, as the specific research will have certain requirements on the software, extending their feature sets, performance, accuracy etc. through dynamic push-pull effects. Conceptualising these as a set of shared core technologies will also have positive effects on their sustainability and interoperability. Also, many European languages other than English are heavily under-resourced, i.e., there are almost no resources or technologies available.

The European academic and industrial technology community is fully aware of the need for sharing resources such as language data, tools and core technology components as a basis for the successful development, implementation and continuous improvement of the Multilingual Services and Applications. Initiatives such as FLaReNet and CLARIN have prepared the ground for a culture of sharing. Services such as META-NET's open resource exchange infrastructure, META-SHARE, can provide the technological platform as well as legal and organisational schemes. This effort will revolve around the following axes: Infrastructure; Coverage, Quality, Adequacy; Language Resources Acquisition; Openness; Interoperability.

[20] http://babelnet.org.

[21] http://linguistic-lod.org/llod-cloud.

5. Research Themes

The four suggested priority research themes are meant to support and further improve the Multilingual Services and Multilingual Applications that will enable the Multilingual Digital Single Market. The suggested activities subsume basic and applied research. In the following we present several concrete ideas, suggestions and indicative examples to illustrate how the research themes are able to drive research and innovation for the Multilingual DSM. Independent of the concrete set of themes, it is important to note that all themes need to be tightly intertwined, making use of one another in different application scenarios, especially so when research results, i.e., technologies, are combined into services and applications.

Multilingual technologies must be at the core of the services and applications for the Multilingual DSM. One of the research themes must tackle high-quality machine translation, including human translation. It needs to provide research results, algorithms, approaches, services, and scientific output that can be directly transformed into generic and specialised services for reliable spoken and written translation among all European and major non-European languages. A second theme must handle crosslingual and multilingual big data analytics of written and/or spoken language data, to provide novel solutions for understanding and dialogue within and across communities of citizens, customers, clients and consumers. This theme needs to include, among others, research scenarios for multilingual sentiment analysis, opinion mining, fact mining, rumour and trend detection, information and relation extraction as well as components that construct semantics for linguistic analyses – taking into account the multitude of established and emerging online text types and genres. A third theme must concentrate on aspects such as conversational technologies, dialogue systems, and natural language interfaces so as to intensify research on interactive spoken language interfaces covering all European languages. Especially with regard to the Internet of Things, and trends such as wearables and Smart Manufacturing (Industrie 4.0), where a very high demand for spoken language interfaces can already now be predicted for the near future. The fourth theme must tackle the increasingly important topic of meaning, semantics, knowledge and data by providing an umbrella for aligning and harmonising all research activities around monolingual, crosslingual and multilingual resources, data sets, repositories, knowledge bases and knowledge graphs that are needed as background knowledge for all advanced language processing components – from machine translation to text analytics to speech interfaces. This theme must take into account more general repositories such as Linked Open Data sets, Wikidata and Wikipedia, multiple different ontologies, OpenStreetMap, DBPedia, but also more research-oriented resources such as Yago, WordNet and BabelNet. Existing and emerging resources need to be consolidated, rendered interoperable, aligned and enriched with multilingual information. Research also needs to work on new approaches for extracting information and knowledge from unstructured text documents and feeding it back into the general knowledge repository. We also need tools for cleaning up data, as well as mechanisms that can aggregate, summarise and repurpose content. For all applications that interact with data, the regulation of intellectual property rights is an issue that needs to be resolved as soon as possible. The web is a global space, and Europe has to find a legal approach that supports both local research, development and innovation while fostering global competitiveness. The key recognition that meaning derives from knowledge also supports a recognition that knowledge is contextual, and users must be taken into account in a way that preserves privacy, retains user control and affords transparent protection of user data.

5.1. Research Theme: Crosslingual Big Data Language Analytics

The central goal behind this theme is to research and to design more precise and more robust language technologies for analysing linguistic Big Data content, not only written text data but also spoken language, in multiple languages with a very broad coverage. In this description we further conceptualise and motivate this research theme with the goal of providing methods, services and applications for improving effectiveness and efficiency of decision-making in business and society by exploiting the digital content of the web.

The research area sketched in this section will change how businesses adapt and communicate with their customers. It will increase transparency in decision-making processes, e.g., in politics and at the

same time give more power to the citizen. As a byproduct, the citizens are encouraged to become better informed in order to make use of their right to participate in a reasonable way. Powerful language analytics will help European companies to optimise marketing strategies or foresee certain developments by extrapolating on the basis of current trends. Leveraging social intelligence for informed decision-making is recognised as crucial in a wide range of contexts and scenarios. Organisations will better understand the needs, opinions, experiences, communication patterns, etc. of their actual and potential customers (trend detection, communication and marketing optimisation). Companies will be able to exploit the knowledge and expertise of their huge and diverse workforce, the wisdom of their own crowds. Companies will be able to adapt to new geographical and cultural contexts. Political decision makers will be able to analyse public deliberation and opinion formation processes in order to react swiftly to ongoing debates or unforeseen events. Citizens will be able to access validated, non-contradictory, multicultural, multilingual and – from multiple political perspectives – information, which will reduce instability and insecurity in Europe. Citizens and customers get the opportunity and information to participate and influence political, economic and strategic decisions of governments and companies, leading to more transparency of decision processes.

These research activities will provide technological support for new forms of issue-based, knowledge-enhanced and solution-centred participatory democracy involving large numbers of expert and non-expert stakeholders distributed over large areas, using multiple languages. The resulting technologies will also be applicable to smaller groups and also to interpersonal communication as well. The research will have a big influence on the Big Data challenge and how we will make sense of huge amounts of data in the years to come.

5.1.1. Novel Research Approaches and Targeted Breakthroughs

Needed to address the Big Data challenge are language technologies that can map large, heterogeneous, and, to a large extent, unstructured volumes of content to actionable representations that support analytics and decision making tasks. Such mappings can range from the relatively shallow to the relatively deep, encompassing, e.g., coarse-grained topic and event-based classification at the document or paragraph/segment level or the identification of named entities, as well as in-depth syntactic, semantic and rhetorical analysis at the level of individual sentences and beyond (paragraph, chapter, text, discourse) or the resolution of co-reference or modality cues within and across sentences.

Technologies such as, e.g., information extraction, entity linking, content validation, reasoning and summarisation have to be made interoperable with knowledge representation and Linked Data as well as Semantic Web methods, e.g., ontological engineering. Drawing expertise from related areas such as knowledge management, information science, or social sciences is a prerequisite to meet the challenge. The research activities should target the bottleneck of knowledge engineering and knowledge acquisition by:

- Semantification of the web: bridging between the semantic islands and the traditional web containing unstructured data.
- Integration of textual and multimedia data with social network and social media data on dimensions including semantics, context, location and temporal; this integration needs to be made possible with regard to typical data representations, i.e., big, heterogeneous, distributed, user-tagged, user-generated, multimodal; the representation has to be lightweight and augmented with semantics including the extraction of semantic representations and transforming them into representations for reasoning and inferencing.
- Aligning and making comparable as well as interoperable different types and genres of content (e.g., news, social media, blogs, academic texts, archives etc.).
- The methods need to be able to operate not only on the actual linguistic data (both spoken language and written texts) but also take into account available metadata, arbitrary markup and other annotations as well as multimedia data (including video, images, audio).
- Among the specific targeted breakthroughs are the following: detecting and monitoring opinions, demands, needs and economic as well as social issues; detecting diversity of views, biases along different dimensions (e.g., demographic) including temporal (evolution of

opinions); support for decision makers and communication participants; problem mining and problem solving; support of collective deliberation and collective knowledge accumulation; vastly improved approaches to sentiment detection and scoring; genre-aware text and language processing; topic recommendation; understanding content and influence diffusion in open and closed social media (identifying drivers of opinion spreading).

- The processing of vast amounts of content also makes it necessary to increase research in the retrieval and summarisation of heterogeneous content sources, e.g., passage retrieval to support question-answering tasks, query rewriting methods and multi-document summarisation with both shallow and deep techniques (including multimodal).
- As a complement, novel approaches are needed with regard to text generation, especially the generation of written and spoken language content based on extracted knowledge (semantic story telling) or based on existing data sets (data-to-text).
- Of specific importance are sophisticated methods for topic and event detection that are tightly integrated with the Semantic Web and Linked Open Data, especially multimodal clustering approaches based on heterogeneous features; automatic and user-driven clustering; clustering and classification based on semantic representations; event detection in multimedia content by exploiting semantic and textual features from speech recognition and captions, as well as visual and motion information.
- In terms of decision support techniques, research is to be intensified on semantic reasoning (rule based, fuzzy, backward and forward chaining) that operates on semantically integrated data, supervised models trained with a variance of features (e.g., concepts, name entities, n-grams, contextual chacteristics, sentiment), visual analytics.
- With regard to the validation and gathering of provenance information, new methods are needed for the detection of fake content, identification of contradictory facts and hidden relations including repeated or similar facts along the spatiotemporal axis.

5.1.2. Solution and Realisation

Solutions should be assembled from a repository of generic monolingual and cross-lingual language technologies, packaging methods in robust, scalable, interoperable, and adaptable components that can be deployed across tasks and projects, as well as across languages where applicable (e.g., when the implementation of a data-driven technique can be trained for individual languages). These need to be combined with approaches that can aggregate data to support decision making and develop new access metaphors and task-specific visualisations. By robust we mean technologically mature, engineered and scalable solutions that can perform high-throughput analysis of web data at different levels of depth and granularity in line with the application requirements. They should be able to work with heterogeneous sources, ranging from unstructured to structured.

To accomplish interoperability we suggest a semantic bias in the choice and design of interface representations: to the highest degree possible, the output (and at deeper levels of analysis also input) specifications of component technologies should be interpretable semantically, both in relation to natural language semantics (lexical, propositional, referential) and extralinguistic semantics (e.g., taxonomic world or domain knowledge). For example, grammatical analysis should make available a sufficiently abstract, normalised, and detailed output, so that downstream processing can be accomplished without further recourse to knowledge about syntax. Event extraction or fine-grained, utterance-level opinion mining should operate in terms of formally interpretable representations that support notions of entailment and inference.

Our adaptability requirement on component technologies addresses the inherent heterogeneity of information sources and communication channels to be processed. Even in terms of monolingual analysis only, linguistic variation across genres (ranging from carefully edited, formal publications to spontaneous and informal social media channels) and domains (as in subject matters) often calls for technology adaptation, where even relatively mature basic technologies may need to be customised or re-trained to deliver satisfactory performance. Further taking into account variation across downstream tasks, big data language processing typically calls for different parameterisations and trade-offs (e.g., in terms of computational cost vs. breadth and depth of analysis) than an interactive

self-help dialogue scenario. For these reasons, relevant trade-offs need to be documented empirically, and component technologies accompanied with methods and tools for adaptation and cost-efficient re-training, preferably in semi- and unsupervised settings.

The solutions needed include high-throughput, big data language analytics that can process multiple multimodal sources, ranging from unstructured to completely structured, at different levels of granularity and depth by allowing to trade-off depth for efficiency as required; extraction of knowledge and semantic integration of social content with sensory data and mobile devices; detection and prediction of events and trends from content and social media networks; technologies for decision support, collective deliberation and e-participation; a large public discussion platform for Europe-wide deliberation on pressing issues such as energy policies, financial system, migration, natural disasters, etc.; mining e-participation content for recommendations, summarisation and proactive engagement of less active parts of population; visualisation of social intelligence-related data and processes for decision support (for politicians, health providers, journalists, manufacturers, entrepreneurs, or citizens).

5.2. Research Theme: High-Quality Machine Translation

The main reason why High-Quality Machine Translation (HQMT) has not been systematically addressed yet seems to be the Zipf'ian distribution of issues in MT: some improvements, the low-hanging fruit, can be harvested with moderate effort in a limited amount of time. Many more resources and a more fundamental, novel scientific approach are needed for significant and substantial improvements that cover the phenomena and problems that make up the long tail. This is a severe obstacle, in particular for individual research centres and SMEs given their limited resources and planning timeframes.

5.2.1. Novel Research Approaches and Targeted Breakthroughs

Although recent progress has already led to new applications of MT technology, radically novel and radically different approaches are needed to accomplish the ambitious goal of this research, i.e., a genuine quality breakthrough. Among these new research approaches are a stronger focus on producing high-quality, publishable outbound translations, needed for the success of MT in the language industry.[22] Research needs to systematically concentrate on the barriers that still prohibit high-quality translations. For this, a fully implemented, unified, dynamic, weighted, and multidimensional quality assessment model with task and language profiling needs to be devised and adopted by the whole research community. This also includes improved automatic quality estimations for given task specifications and the inclusion of translation professionals in the research and innovation process. Vice versa, human translation needs to be enhanced with ergonomic, computer-supported work environments and multilingual text authoring. The recent breakthroughs in neural MT need to be improved through additional statistical models that extract more dependencies from the data. In addition, a semantic translation paradigm is needed by extending statistical translation with semantic data such as linked open data, ontologies including semantic models of processes and textual inference models. We also want to put a stronger emphasis on the properties of individual languages, especially through the exploitation of strong monolingual analysis as well as generation methods and resources. Furthermore, we want to intensify research on modular combinations of specialised analysis, generation and transfer models, permitting accommodation of registers and styles (including user-generated content) and also enabling translation within a language (e.g., between specialists and laypersons).

The expected breakthroughs will include high-quality text translation and reliable real-time speech translation for all official European languages as well as regional and minority languages; a modular analysis-transfer-generation translation technology that facilitates reuse and constant improvement of (statistical and knowledge-driven) modules; automatic subtitling and voiceover of films and

[22] As opposed to the dominant information gisting paradigm which has been pushed by (US) intelligence interests and is of course also relevant for many applications where approximate translations are sufficient or no translations could be provided otherwise.

multimedia applications in selected domains, such as public service, sports events, and other applications; always-correct translation for critical subdomains.

5.2.2. Solution and Realisation

Cooperation with translation professionals. A close cooperation of language technology research and professional language service experts is foreseen. The knowledge of translators and post-editors will provide judgements and corrections for insights towards a more analytical and systematic approach of quality boundaries and data for bootstrapping new methods. The cooperation scheme will be fruitful since language service professionals or experts in translation studies will also be the first test users analytically monitored by the evaluation schemes. This symbiosis will lead to a better interplay between research and innovation.

Novel quality metrics and human annotation. The improvements needed for HQMT have to be based on novel, reliable and informative quality measures since common measures such as, e.g., BLEU or TER, may incorrectly punish perfectly fine translations, if they differ from a given reference translation. Currently, the only way of assessing translation quality involves manual work such as post-editing or error annotations. This data is needed for system development and as test cases for evaluating the performance of new models using advanced diagnostic tools. The medium term goal is to automate novel metrics as far as possible including sampling functionality, to incorporate feedback from research systems and to develop datasets for new metrics and best practices.

Exploiting human annotations for improving models. Error annotations and post-edits on industry-derived MT output is to be analysed to determine to what degree annotations and edits can be predicted or automated. Established string-based matching metrics will be extended with syntactic and semantic information from parsing or role labelling. The class of features that correlates with the annotations of human translators will be used to inform both translation and quality estimation models and help researchers to make their development cycles more targeted and focussed. MT will be improved both system-internally and externally: At upstream level, source sentences will be automatically adapted to increase their translatability. At downstream level, target sentences will be automatically corrected accounting for their expected final use (e.g., gisting, publishable translation). At system level, the acquired correction rules will be used to project knowledge onto the core MT system components. This will enable a continuous self-learning framework where the selection of proper model extension or updating strategies will be driven by penalisation and rewarding criteria.

Platform for MT research and development. The procedures outlined above pertain to both MT development in research and production. It should be tested and further developed into more standardised pipelines. A large-scale evaluation infrastructure, structured to areas, applications, and languages is to be designed and implemented for the resource and evaluation demands of large-scale collaborative research. An initial inventory of tools and resources as well as extensive experience in shared tasks and evaluation has been obtained in several EU-funded projects. Together with LSPs, a common service layer supporting research workflows on HQMT must be established. As customer data is needed for realistic development and evaluation, IPR and legal issues must be taken into account. The platforms to be built include trusted service clouds, workbenches for translators and translation workflows.

5.3. Research Theme: Meaning, Semantics, Knowledge

While Machine Translation is a key technology for the Multilingual Digital Single Market, other technologies are needed to tailor engagement between companies and their customers to the domain being addressed. Customers should be able to search for user-generated content (UGC) on a specific product or service regardless of the language in which it was posted. Image, video and audio postings on products should be tagged, summarised, discoverable and accessible to users in any other language. Customer profiles should be built in their native language so the personalisation of engagement can be automated in that language, while still providing market intelligence in the vendor's native language. To successfully tailor such cross- and multilingual customer experiences, companies must monitor and analyse UGC on social media, blogs, forums and product review sites and react continuously with well

targeted customer engagement. The effectiveness of Language Technologies is, however, limited by the distance between the linguistic data available to train them and the content they must process when deployed in a specific application. This is especially problematic for SMEs. Small companies succeed by excelling in a specific niche where they must engage skilfully with their customers using and understanding the terms and language patterns specific to that niche. One-size-fits-all language technologies, such as unrestricted machine translation, will fail to meet the language needs of specialised SMEs. Small companies, however, typically lack the knowledge or capacity to assemble their own linguistic data assets to tailor language technology to their needs. Without tailored language technology support, though, SMEs will not be able to make use of the DSM because of the language barriers to bidirectional customer engagement. Linguistic Linked Data is already proving a scaleable source of massively multilingual open language resources for LT services. Research is needed into tools and techniques to integrate the lifecycle management of linguistic data into technologies that apply to the specific niches of online discourse that SMEs must use. SMEs must be empowered with cheap and easy tools to assemble, deploy and refine micro-domains for linguistic and semantic resources that can be used across different LT components they employ.

Data has been referred to as the new oil of the digital economy. However, crude oil is useless unless it is refined. The same holds for multilingual data. If data is not linked to other data it can only be used in isolation, rather than in context. If data is not analysed further, no insights can be generated. If data is not verified nor the provenance of data tracked, it cannot be trusted. If the licensing terms under which data is provided are not known, then it cannot be exploited appropriately. Linking, deeper analysis, verification and validation, provenance attribution and clear indication of licensing terms are crucial to create an ecosystem in which multilingual data can be safely and meaningfully exploited in data value chains that generate insights.

5.3.1. Novel Research Approaches and Targeted Breakthroughs

The main dimensions that need to be prioritised are the following:

Linking: Only if knowledge repositories, knowledge graphs and data sets are linked across sources can they be exploited in context, making more of the single data set compared to using it in isolation. Linking is crucial to exploit data, investments in new methodologies for knowledge linking are needed. As the amount of knowledge and data grows, it will become harder and harder to find the item that is most appropriate to solve a particular task. We need to create an ecosystem that fosters knowledge and data discovery. We need to create an ecosystem for multilingual data and knowledge, in which links are first class, value-adding objects and tools are available to manage the relevance, authoritativeness and quality of links.

Generating insights from unstructured data: Data is often in unstructured form, such that it cannot be directly exploited in applications or to generate insights. Robust, efficient and scalable techniques for refining unstructured data in such a way that it can be transformed to make it usable are needed. Human language technologies and NLP methods play a crucial role and need to be extended in terms of coverage, robustness and scalability.

Trust and Usability: For knowledge and data to be exploited in applications, trust in the data is key. It involves knowing where the data comes from and who generated it, but also knowing which permissions, prohibitions and implications come with the data to ensure compliance with the terms of use with the data. Provenance and licensing information must remain attached to data over the whole data lifecycle (creation, use, derivation, modification).

Privacy and Data Protection: An ecosystem of knowledge and data needs to respect the right of people for privacy and empower them to decide who can use their personal data for which purpose. We need an ecosystem in which data use is made transparent so that users are aware of the implications of providing data to a certain entity and they are empowered to retract their data at any point. There are massive new challenges in how users understand and control how their spoken or written utterances are used.

Universal access to data commons and public services across languages: The emerging data commons cannot remain exclusively exploited by experts or companies with huge infrastructures and resources. Instead, we need to make sure that also the public at large can benefit from data by simplifying access and use across languages.

Access to information and services without borders: We cannot afford that access to data and services stops at the borders of countries due to language barriers. We need to substantially invest into the cross-border flow of data and availability of public and commercial services but also in homogenisation and consistency of services across borders and languages. This requires the integration of language technologies and localisation into knowledge, semantic and linked data technologies, in particular through the use of standards.

If Europe does not substantially invest in the above and several other closely related fields, it will most certainly fall behind other international competitors. Europe has failed in the past to invest in search technology and has no alternatives of its own to offer to the market leaders in the US and Asia. This is a key failure as it implies that big players from other countries are deciding what European citizens find online and with what level of privacy. This is a threat to the free flow of information that runs counter to the free and independent availability of information that is required to strengthen democracy that is key to the European tradition.

5.3.2. Solution and Realisation

The main bottleneck of the Semantic Web remains the problem of knowledge acquisition. The intellectual construction of domain models turned out to be an extremely demanding and time-consuming task, requiring well-trained specialists that prepare new ontologies from scratch or base their work on existing taxonomies, ontologies, or categorisation systems. Information extraction can be used for learning and populating ontologies from unstructured knowledge. Texts and pieces of texts can be annotated with extracted data. These metadata can serve as a bridge between the semantic portions of the web and the traditional web of unstructured data, providing unprecedented levels of contextualised knowledge. For connecting between different media in the multimedia content of the web, some of the needed tasks are annotating pictures, videos, and sound recordings with metadata, interlinking multimedia files with texts, semantic linking and searching in films and video content, and cross-media analytics, including cross-media summarisation.

In the Jeopardy game show, IBM's Watson was able to find correct answers that none of its human competitors could provide, which might lead one, erroneously, to think that the problem of automatic question answering is solved. With clever lookup and selection mechanisms for the extraction of answers, Watson could actually find the right responses without a full analysis of the questions from a huge set of handbooks, decades of news, lexicons, dictionaries, bibles, databases, and the entire Wikipedia. Outside the realm of quiz shows, however, most questions that people might ask cannot be answered by today's technology, even if it has access to the entire web, because they require a level of language and context understanding that is not possible yet with today's technology. Modelling the contexts in which users ask questions must therefore be efficiently indexed against into the increasingly massive body on multilingual knowledge from which answers can be sourced.

Linking knowledge to rich interaction corpora will enable the development of agents which can assist proactively and can make inferences from their own limited knowledge, to enable people to be notified of relevant things faster, and to help people reach understanding of complex situations involving many streams of information. By 2025, we envisage such systems which operate on huge, dynamic, heterogeneous data streams. It will be important to consider issues such as provenance, trust, privacy, data protection, security, and rights. Compliance with applicable standards relating to these matters will have to be designed into the platform from the outset. A key issue for this scenario relates to positive (democracy) and negative (surveillance) aspects of large-scale multimodal knowledge integration and access.

5.4. Research Theme: Conversational Technologies

Conversational agents and interactive dialogue systems that enable voice-controlled interfaces with multilingual capabilities will play a crucial role for the multilingual Digital Single Market. This not only relates to connected devices (Internet of Things) but also to apps such as chat bots. More details on this research topic are provided in the roadmap by the CITIA Alliance.

The overall goal of the Conversational Technologies community is to develop and make operational socially aware, multilingual systems that support users interacting with their environment, including human-computer, human-agent (or robot), and computer-mediated human-human interaction. Systems must be able to communicate, exchange information and understand other agents' intentions. They must be able to adapt to the user's needs and environment and have the capacity to learn from all interactions and sources of information.

The ideal interactive system can interact naturally with humans, in any language and modality. It can adapt and be personalised, including special needs (for the visual, hearing, or motor impaired), affections, or language proficiencies. It can recognise and generate speech incrementally and fluently. It can learn, personalise itself and forget. It can assist in language training and education. It recognises people's identity, and their gender, language or accent. If the agent is embodied in a robot, it can move, manipulate objects, and interact with people.

This research theme includes several core components: Interacting naturally with users in an implicit (proactive) or explicit (spoken dialogue and/or gesture) manner based on robust analysis of human user identity, age, gender, verbal and nonverbal behaviour, and social context; using language in connection with other communication modalities (visual, tactile, haptic); exhibiting robust performance; interacting naturally with and in groups (in social networks, with humans or artificial agents/robots); exhibiting multilingual proficiency (translation, interpretation in meetings and videoconferencing, cross-lingual information access); referring to written support (transcription, close captioning, reading machines, ebooks); providing access to knowledge (answers to questions, shared knowledge in discussion); providing personalised training; dialogue systems evaluation needs more research on the choice of adequate metrics and protocols. The multilingual dimension that is targeted implies the availability of language resources and technology evaluation for all languages.

5.4.1. Novel Research Approaches and Targeted Breakthroughs

The development of conversational technologies requires several research breakthroughs. With regard to speech recognition, accuracy (open vocabulary, any speaker) and robustness (noise, cross-talking, distant microphones) have to be improved. Methods for self-assessment, self-adaptation, personalisation, error-recovery, learning and forgetting information, and also for moving from recognition to understanding have to be developed. In speech synthesis, voices have to be made more natural and expressive, parameters have to be included for meaning, style and emotion. They also have to be equipped with methods for incremental speech, including pauses and hesitations..

As human communication is multimodal (including speech, facial expressions, body gestures, postures, etc.), crossmodal and fleximodal, generic semantic and pragmatic models of human communication have to be developed. These have to be context-aware to model situational interdependencies between context and modalities for arriving at robust communication analysis. They have to be able to detect and recover interactively from mistakes, learning continuously and incrementally. To be able to design technologies, adequate semantically and pragmatically annotated language and multimodal resources have to be produced.

A common push has to be made towards more natural dialogue. This includes, among others, the recognition and production of paralinguistics (prosody, visual cues, emotion) and a better understanding of socioemotional functions of communicative behaviour, including group dynamics, reputation and relationship. In addition, more natural dialogue needs more advanced dialogue models that are proactive (not only reactive), that are able to detect that recognised speech is intended as a machine command, they have to be able to interpret silence as well as direct and indirect speech acts (including lies and humour). Another prerequisite for more natural dialogue is the ability of the system

to personalise itself to the user's preferences. The system has to operate in a transparent way and be able to participate in multi-party conversations and make use of other sensory data (GPS, RFID, cameras etc.).

The multilingual assistant should also be able to do translation in human-human interaction and to deal with different languages, accents and dialects effectively. Systems developed should also cover at least all official languages of the EU and several regional languages.

5.4.2. Solution and Realisation

The scientific state-of-the-art is at a stage that finally allows tackling the development of robust conversational technologies. Progress in machine learning, including adaptation, unsupervised learning from data streams, continuous learning, and transfer learning makes it possible automatically to learn certain capabilities. Existing language and multimodal resources enable the bootstrapping of systems. Furthermore, there is interdisciplinary progress made in, e.g., social signal processing and also knowledge representation including approaches such as the Semantic Web and Linked Open Data – especially inferences and automatic reasoning are an important prerequisite. Technological advances are continuously being achieved in the vision-based human behaviour analysis and synthesis fields. Ubiquitous technologies are now widely available. User-centric approaches have been largely studied and crowd sourcing is used more and more. Quantitative and objective language technology and human behaviour understanding technology evaluations, allowing for assessing a technological readiness level, are carried out more widely and language resources and publicly-available annotated recordings of human spontaneous behaviour are now available. However, there are prohibitive factors. Evaluation is still limited and not conducted for all languages. There is limited availability of language resources. Publicly available recordings of spontaneous human behaviour are sparse, especially when it comes to continuous synchronised observations of multiparty interactions. Limited progress of the technology for automatic understanding of social behaviour like rapport, empathy, envy, conflict, etc., is mainly attributed to this lack of suitable resources. In addition, we still have limited knowledge of human language and human behaviour perception processes. Automated systems often face theoretical and technological complexity of modelling and handling these processes correctly.

6. Horizontal Topics

In this chapter we briefly discuss several horizontal aspects of the MLV Programme.

6.1. Standardisation and Interoperability

Especially for the successful design, implementation, deployment and continuous improvement of the services and platforms, efforts for ensuring the interoperability of methods and services need to be intensified by significantly boosting standardisation activities – not only as an afterthought but already during the research, development and innovation phase of the implementation of the MLV Programme. In order to provide a few concrete examples, we have added suggestions for hands-on standardisation topics in the definition of the MLV Programme. These topics are to be embedded in innovation actions to avoid a gap between standardisation and real world use cases.

6.2. Business Models and Ecosystems

We anticipate an intensified discussion of business models and ecosystems are language technologies, especially with regard to Multilingual Services and Multilingual Applications (see Chapter 2). A set of interconnected Coordination and Support Actions should take care of finding synergies among the different subfields and tie these discussions together in order to provide projections and best practice examples. This approach is in line with two concrete goals formulated by some of our stakeholders, listed in the following:

- **2020:** Enable localisation industry to explore new business models, beyond translation, and contributing, e.g., to marketing of digital content

- **2020:** Build a connection between creators, distributors and consumers of public sector information (PSI), to allow for feedback on the usefulness of public data sets in new business models

6.3. Language Policies and Public Procurement

Technology progress would be even more efficient and effective if the recommended MLV Programme could be accompanied by appropriate supportive policy making in several areas. One of these areas is multilingualism. Overcoming language barriers can greatly influence the future of the EU and the whole planet. Solutions for better communication and for access to content in the native languages of the users would not only enable the multilingual Digital Single Market, it would reaffirm the role of the EC to serve the needs of the EU citizens. A substantial connection to the infrastructural programme Connecting Europe Facility (CEF) could help to speed up the transfer of research results to badly needed services for the European economy and public. At the same time, use cases should cover areas where the European societal needs massively overlap with business opportunities.

Language policies supporting multilingualism can create a tangible boost for technology development. Some of the best results in Machine Translation have been achieved in Catalonia, where legislation supporting the use of the Catalan language has created an increased demand for automatic translation.

Numerous US breakthroughs in IT that have subsequently led to successful products of great economic impact were only achieved by a combination of systematic long-term research support and public procurement. Many types of aircraft or the autonomous land vehicle would not have seen the light of day without massive government support – even the internet or the speech technology behind Apple Siri benefited largely from sequences of DARPA programmes often followed by government contracts procuring earlier versions of the technology for military or civilian use by the public sector.

The search for originality on the side of the public research funding bodies and their constant trial-and-error search for new topics that might finally help the European IT industry to be in time with their innovations have often caused the premature abortion of promising developments, whose preliminary results were more than once taken up by research centres and enterprises in the US. An example in language technology is the progress in statistical machine translation. Much of the groundwork laid in the German government-sponsored project Verbmobil (1993–2000) was later taken up by DARPA research and commercial systems – including Google Translate.

Today, outdated legislation and restrictive interpretation of existing law hinder the effective use of many valuable data collections such as, for example, several national corpora. The research community urgently needs the help of European and national policy makers for modes of use of these data that would boost technology development without infringing on the economic interests of authors and publishers.

2018: In order to drive technology evolution with public funding to a stage of maturity where first sample solutions can deliver visible benefits to the European citizens and where the private sector can take up technologies to then develop a wide range of more sophisticated profitable applications, we strongly recommend a combination of

1. language policies supporting the status of European languages in the public sector,
2. procurement of solution development by European public administrations,
3. long-term systematic research efforts with the goal to realise badly needed pre-competitive basic services.

European policy making should also speed up technology evolution by helping the research community to gain affordable and less restrictive access to text and speech data repositories, especially to data that have been collected with public support for scientific and cultural purposes.

6.4. Copyright and Data Protection

2019: Research and innovation in language technology depends on language data the way climate research depends on weather data or economic studies depend on financial data. Results derived in language technology research from the analysis of large amounts of texts in areas like machine translation, text mining or text analytics such as statistical models or abstract representations do not interfere with the copyright holders' rights to publish, republish, modify, translate and otherwise make available the texts in order for someone else to read them as a document, piece of art, etc. Still, traditional copyright and half-hearted exceptions for research are experienced as huge obstacles for research and innovation by the European research community – the EU Fair Use principle can be applied in some cases but, in general, more needs to be done. These obstacles come with a threat of severe economic consequences: academic and industrial researchers – already a sparse resource – may leave Europe to pursue their goals in other continents, technology leadership may migrate to the US or Asia, immense opportunities of growth are lost. We are happy that the EC is taking the next steps towards the important and urgent goal of a reform of European copyright law.

6.5. Open Source

While language technology-based industry solutions target an agile high-tech industry, many fields still appear to be dominated by expensive and slow-moving monolithic as well as proprietary software that makes it especially hard for many SMEs to compete with developments. At the same time other areas have shown that massive collaboration in open-source-projects can lead to impressive and future-proof software such as operating systems (e.g., Linux) or CMS platforms (e.g., Drupal).

Still, open source projects usually do not run by themselves. They require well conceived forms of organisation fitting the respective community and type of project. Therefore, these developments need to be supported by platforms and funding schemes in their own right.

While we do not want to play off proprietary against open-source software, we do want to support the development of the latter for the language industry. In fact, many tools and standards already exist in the industry and in language technology research, i.e., open source development is the normal case. Existing tools are often not mature enough and lack plans for maintenance so that they are only of limited usefulness for the industry and public services.

6.6. Related Areas, Applications and Societal Challenges

The applications, solutions, services, infrastructures and tangible outcomes of the MLV Programme will not only create the multilingual Digital Single Market. Several closely related areas and applications as well as societal challenges will benefit from them as well.

Most evident is the complementary connection to the BDV cPPP in terms of technologies for multilingual big data analytics and cross-lingual data value chains. There is also a close relationship to interactive and multilingual spoken language interfaces and robots (especially the SPARC Robotics PPP), connected machines (Advanced Manufacturing, Industrie 4.0), Inclusion, E-Learning as well as generic connected devices (Internet of Things, Web of Things). The relationship between multilingual technologies and ecommerce applications is so evident and of such vital importance that we also mention this area, as well as the emerging trend to Smart Cities and Smart Services.

The importance of the languages in our European society has never been in the focus of attention as compared to other highly multilingual societies like South Africa or India where language borders hinder exchange and communication *within* a state. According to the principles of the UN-endorsed World Summit on the Information Society, the "Information Society should be founded on and stimulate respect for cultural identity, cultural and linguistic diversity." Recent scientific work has

shown that even our moral decisions are influenced by whether we are speaking our mother tongue or a foreign language.[23]

In fact, the technology solutions detailed in the next chapter address many of the societal challenges specifically to be taken into account by activities under the framework of Horizon 2020.[24] The following list provides several examples:

- Health, demographic change and wellbeing (can be addressed by Adaptable interfaces for all, E-Health, and E-Learning solutions);
- Food security, sustainable agriculture and forestry, marine and maritime and inland water research, and the bioeconomy (can be addressed by the Digital Translation Centre solution);
- Secure, clean and efficient energy (can be addressed by the E-Participation solution);
- Smart, green and integrated transport (can be addressed by the Adaptable interfaces for all solution);
- Climate action, environment, resource efficiency and raw materials (can be addressed by Digital Translation Centre solution);
- Europe in a changing world – inclusive, innovative and reflective societies (can be addressed by Adaptable interfaces for all, E-Learning, E-Participation solutions);
- Secure societies – protecting freedom and security of Europe and its citizens (can be addressed by Adaptable interfaces for all solution).

7. Conclusions

7.1. Expected Economic Impact

The EC predicts that the transition to the integrated Digital Single Market will deliver up to €400 billion in economic growth by 2020. However, this ambitious goal – in fact, even more – can only be reached if the language factor is taken into account. If customers are still hampered by language, online commerce will remain confined to fragmented markets, which are defined by language silos. Approximately 60% of individuals in non-Anglophone countries seldom or never make online purchases from English-language sites; the number willing to purchase from sites in non-native languages other than English is much, much lower. As a result, no language can address 20% or more of the DSM.

European SMEs are an integral and vital component of the DSM. However, only 15% of them sell online – and of that 15%, fewer than half do so across borders. SMEs that sell their products and services internationally exhibit 7% job growth and 26% innovate in their offering – compared to a job growth of 1% and 8% innovation for SMEs that do not sell their products and services internationally. Only if Europe accepts the multilingual challenge and decides to design and to implement research and innovation driven technological solutions as well as a service infrastructure with the goal of overcoming language barriers, can the economic benefits of the DSM be achieved. Enabling and empowering European SMEs easily to use language technologies to grow their business online across many languages is key to boosting their levels of innovation and jobs creation.

If the MLV Programme specified in this Strategic Agenda is fully realised, we expect the economic growth by 2020 to be much higher than the predicted €400 billion since, crucially, we will have successfully enabled many European SMEs to sell online on the *multilingual* Digital Single Market, substantially multiplying their reach. Furthermore, we expect the creation of tens of thousands of sustainable new jobs in the medium to long-term. The growth would not stop at the borders of Europe: if the strategic programme is successful, Europe could offer the developed solutions to other multilingual societies, for example, to adapt and to export certain parts of the MLV Programme to India or South Africa.

---

[23] A. Costa, A. Foucart, S. Hayakawa, M. Aparici, J. Apesteguia, J. Heafner, B. Keysar (2014): "Your Morals Depend on Language", PLOS One, http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0094842.

[24] European Commission (2014): Horizon 2020, The EU Framework Programme for Research and Innovation, Societal Challenges, http://ec.europa.eu/programmes/horizon2020/en/h2020-section/societal-challenges.

The European DSM today would account for approximately 25% of global economic potential. However, if Europe overcame the language barriers that hamper intra-European trading, it would also remove barriers to international trade that keep European SMEs from achieving their full economic potential by penetrating markets in other continents beyond our own. Addressing the official and major regional languages of Europe would open access to over 50% of the world's online potential and 73% of the world online market in economic terms, amounting to an online market of approximately €25 trillion (sic!) in 2013. *The global potential for European businesses exceeds the continent-internal opportunities from the DSM by orders of magnitude.*

7.2. Potential Funding Sources

We suggest setting up, under the umbrella of the MLV Programme, a coordinated initiative both on the international (EC/EU) and national level (Member States, Associated Countries, regions), including research centres as well as small, medium and large enterprises who work on or with language technologies and other stakeholders, especially user companies.

The European Union could support the MLV Programme especially through dedicated activities in upcoming Horizon 2020 calls (2018–2020) and through Connecting Europe Facility (CEF). Horizon 2020 Research Actions are compatible to our planned activities in the area of research, while Horizon 2020 Research and Innovation Actions as well as Coordination and Support Actions are needed for the actual innovation and deployment activities. Highly innovative activities with a major commercial impact are needed for the Application Areas – especially here, the European language technology industry will participate (most of these companies are SMEs). Through CEF, deployment and innovation actions could be funded, especially with regard to public online services. Furthermore, there are horizontal programmes such as Horizon 2020 Widespread/Teaming that could boost the knowledge and technology transfer between countries that already have excellent research and innovation hubs in language technology and those that do not; the goal would be to enable the less innovative countries to develop technologies for their respective languages. Similar programmes to boost SMEs exist.

On the national and regional levels, the respective local funding agencies could provide resources, especially to support the development of technologies for their respective national or regional languages. There are also dedicated programmes for supporting national and regional companies becoming more innovative.

Critically, public procurement can play a decisive role in this strategic programme: if the European Union is willing to invest in the development of multilingual technologies made *in* Europe and apply them *for* Europe, the EU itself would be the perfect reference user of such technologies, setting an example for national or regional governments.

7.3. Next Steps

This document represents the second evolutionary version after an initial suggestion that was prepared by the European language technology community. This initial version (Version 0.5) was publicly unveiled at META-FORUM 2015 and the Riga Summit 2015 on the Multilingual Digital Single Market (April 27-29, 2015, in Riga, Latvia).[25] At the Riga Summit 2015, we also initiated the first public consultation phase. Feedback and additional input gathered during the event has flown back into the current version of the strategic agenda (Version 0.9). This current version will be presented at META-FORUM 2016 in Lisbon, Portugal.

As soon as the MLV programme has been discussed with the European Commission and an agreement has been reached, the strategies and roadmaps need to be further aligned, refined and specified in the community in one more public consultation phase, especially concerning the bridge to the Big Data Value Association.

We expect the final version of this document to be available in late 2016.

[25] http://rigasummit2015.eu

Appendix

A. Editorial Team

Representatives from the EU project CRACKER:
Aljoscha Burchardt, Jan Hajic, Georg Rehm, Lucia Specia, Hans Uszkoreit, Josef van Genabith

Representatives from the EU project LT_Observatory:
Gerhard Budin, Steven Krauwer, Vesna Lusicky

Representatives from the Cracking the Language Barrier federation:
Kalina Bontcheva, Steve Renals, Felix Sasaki, Andrejs Vasiljevs

B. History of this Document

Version 0.5: Second version, after a more complex and longer preliminary version; V0.5 presented at META-FORUM 2015 and Riga Summit 2015.

Collected feedback and additional input after Riga Summit 2015.

Version 0.9: Third version, presented at META-FORUM 2016.

Feedback and additional input to be collected at and after META-FORUM 2016.

We have established a link with the editorial team of the BDVA SRIA and will make sure that the two documents will be kept aligned in the future.

Version 1.0: To be presented in September/October 2016.

C. Input Documents

The following documents, roadmaps and presentations have informed the current version of the Strategic Agenda for the Multilingual Digital Single Market.

- Philipp Cimiano (2015): "The LIDER Roadmap in a nutshell", presentation given at the Workshop on multilingual data value chains in the Digital Single Market, 16 January 2015, Brussels, Belgium. https://ec.europa.eu/digital-agenda/en/news/workshop-multilingual-data-value-chains-digital-single-market.
- Gerald Cultot (2015): "eHealth services – multilingual challenges", presentation given at the Workshop on multilingual data value chains in the Digital Single Market, 16 January 2015, Brussels, Belgium. https://ec.europa.eu/digital-agenda/en/news/workshop-multilingual-data-value-chains-digital-single-market.
- Andrew Joscelyne (late 2014): "A Strategic Research and Innovation Agenda for a Conversational European Digital Marketplace" (draft position paper).
- Nils Lenke (2015): "Nuance Inc.", DFKI Tech Day, 30 January 2015, DFKI Saarbrücken, Germany.
- Dave Lewis (2015): "Shopping Across the Language Barrier", presentation given at the Workshop on multilingual data value chains in the Digital Single Market, 16 January 2015, Brussels, Belgium. https://ec.europa.eu/digital-agenda/en/news/workshop-multilingual-data-value-chains-digital-single-market.
- LIDER (10/2014): "Roadmap for the use of Linguistic Linked Data for content analytics"
- META-NET (2013): "Strategic Research Agenda for Multilingual Europe 2020", Georg Rehm and Hans Uszkoreit (eds.), presented by the META Technology Council. Springer.
- MLi (09/2014): "D5.1 – Big and Social Language Data Requirements for the MLi Hub".

- QTLaunchPad (11/2014): "European Quality Translation Research 2015: Ongoing Work and Roadmap".
- Ruben Riestra (2015): "Multilingual data value chains in the Digital Single Market", report presented at the Workshop on multilingual data value chains in the Digital Single Market, 16 January 2015, Brussels, Belgium. https://ec.europa.eu/digital-agenda/en/news/workshop-multilingual-data-value-chains-digital-single-market.
- ROCKIT (10/2014): Roadmap for Conversational Interaction Technologies – Coordination and Support Action, D1.1 Innovation Drivers, future scenarios, and best practice.
- ROCKIT (10/2014): Roadmap for Conversational Interaction Technologies – Coordination and Support Action, D2.1 First Report on Innovation in the ROCKIT Domain.
- ROCKIT (10/2014): Roadmap for Conversational Interaction Technologies – Coordination and Support Action, D3.1 First Report on Research in the ROCKIT Domain.
- ROCKIT (02/2014): Roadmap for Conversational Interaction Technologies – Coordination and Support Action, D4.1 ROCKIT Roadmap Specifications.
- Alan Mas Soro: "Language Technologies for Europe: A way to foster SME internationalization", presentation given at the Workshop on multilingual data value chains in the Digital Single Market, 16 January 2015, Brussels, Belgium. https://ec.europa.eu/digital-agenda/en/news/workshop-multilingual-data-value-chains-digital-single-market.
- Adomas Svirskas (2015): "Pan-European Electronic Document Platform. Open Interoperable Solution for Europe", presentation given at the Workshop on multilingual data value chains in the Digital Single Market, 16 January 2015, Brussels, Belgium. https://ec.europa.eu/digital-agenda/en/news/workshop-multilingual-data-value-chains-digital-single-market.
- Hans Uszkoreit (2014): "European Platform(s) for Machine Translation and other Language Technologies", presentation given at the META-NET Platform Strategy Meeting during the Language Resources and Evaluation Conference (LREC), 26-31 May 2014, Reykjavik, Iceland.
- Xenios Xenophontos (2015): "Online Dispute Resolution Platform – Multilingual challenges", presentation given at the Workshop on multilingual data value chains in the Digital Single Market, 16 January 2015, Brussels, Belgium. https://ec.europa.eu/digital-agenda/en/news/workshop-multilingual-data-value-chains-digital-single-market.
- Sonja Zillner (2015): "cPPP Big Data Value-SRIA", presentation given at the Workshop on multilingual data value chains in the Digital Single Market, 16 January 2015, Brussels, Belgium. https://ec.europa.eu/digital-agenda/en/news/workshop-multilingual-data-value-chains-digital-single-market.

D. Digital Language Extinction in Europe

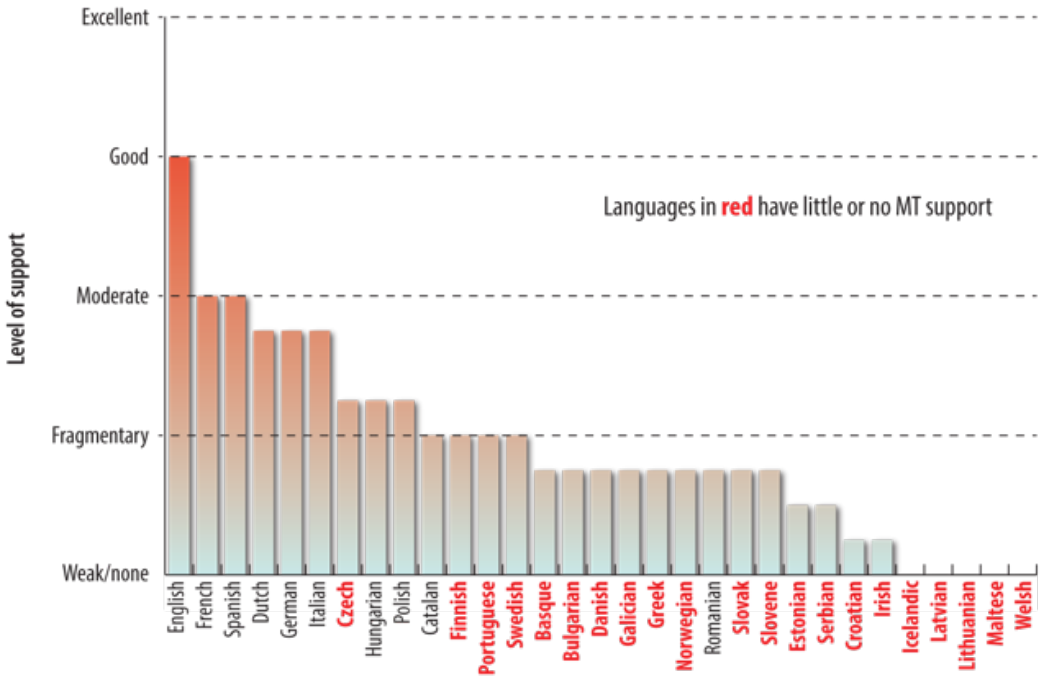
Most European languages are unlikely to survive in the digital age, a study by Europe's leading Language Technology experts warns. Assessing the level of support through language technology for 30 of the more than 60 European languages, we concluded that digital support for 21 of the 30 languages investigated is "non-existent" or "weak" at best. The study "Europe's Languages in the Digital Age" was carried out by META-NET, a European network of excellence that consists of 60 research centres in 34 countries, working on the technological foundations of multilingual Europe.

Europe must take action to prepare its languages for the digital age. They are a precious component of our cultural heritage and, as such, they deserve future-proofing. The META-NET study shows that, in the digital age, multilingual Europe and its linguistic heritage are facing challenges but also many possibilities and opportunities.

The study, prepared by more than 200 experts and documented in 31 volumes of the META-NET White Paper Series (available both online and in print), assessed language technology support for each language in four different areas: automatic translation, speech interaction, text analysis and the availability of language resources. A total of 21 of the 30 languages (70%) were placed in the lowest

category, "support is weak or non-existent" for at least one area by the experts. Several languages, for example, Icelandic, Lithuanian and Maltese, receive this lowest score in all four areas but it must be noted that support for some of the languages with smaller numbers of speakers is slowly increasing since the original publication of the META-NET White Paper Series in 2012. At the other end of the spectrum, while no language was considered to have "excellent support", only English was assessed as having "good support", followed by languages such as Dutch, French, German, Italian and Spanish with "moderate support". Languages such as Basque, Bulgarian, Catalan, Greek, Hungarian and Polish exhibit "fragmentary support", placing them also in the set of high-risk languages.

The white papers and more details are available at http://www.meta-net.eu/whitepapers.



Investment in the following Multilingual Applications and Multilingual Services* will help achieve the Multilingual Digital Single Market

* (including online and public services)

Unified Customer Experience

- Provides a contextualised experience to users (for multi-lingual e-commerce)
- Brings together content, product, customer care, customer relation-ship, discussion fora, help-desks etc.
- Unified digital (eco)system across languages

Voice of the Customer and Voice of the Citizen

- Comprehensive methods for multilingual market research and Europe-wide crosslingual demographics and surveys
- Connects business to customer opinion and politics to citizen opinion – across borders and languages

Digital Translation Centre

- Automatic translation services
- Free (for the citizen) or for a fee (specialised HQ services)
- To and from businesses, govern-ments, customers, citizens, public institutions

Content Curation and Production

- Smart multilingual authoring support
- Multilingual and multimodal report generation, cross-lingual linking, enrichment, and semantification

The editorial team of this Strategic Research and Innovation Agenda (SRIA) can be reached through Dr. Georg Rehm: georg.rehm@dfki.de.

This document has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements No. 645357 (CRACKER) and No. 644583 (LT_Observatory).