This document is part of the Coordination and Support Action CRACKER. This project has received funding from the European Union's Horizon 2020 program for ICT through grant agreement no.: 645357.

Deliverable D5.5

Preliminary joint Strategic Research and Innovation Agenda for the LT/MT field

Authors: Aljoscha Burchardt, Georg Rehm (DFKI)

Dissemination Level: Public

Date: 31 March 2015

Grant agreement no.	645357
Project acronym	CRACKER
Project full title	Cracking the Language Barrier
Type of action	Coordination and Support Action
Coordinator	Dr. Georg Rehm (DFKI)
Start date, duration	1 January 2015, 36 months
Dissemination level	Public
Contractual date of delivery	03/2015
Actual date of delivery	03/2015
Deliverable number	D5.5
Deliverable title	Preliminary joint Strategic Research and Innovation Agenda for the LT/MT field
Type	Report
Status and version	Version 1.0 of this deliverable reports on SRIA version 0.2
Number of pages	96
Contributing partners	DFKI
WP leader	DFKI
Task leader	DFKI
Author(s)	Aljoscha Burchardt, Georg Rehm (DFKI)
Internal reviewers	n.a. (SRIA prepared and reviewed in a community process)
EC project officer	Pierre-Paul Sondag
The partners in CRACKER are:	• Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Germany • Charles University in Prague (CUNI), Czech Republic • Evaluations and Language Resources Distribution Agency (ELDA), France • Fondazione Bruno Kessler (FBK), Italy • Athena Research and Innovation Center in Information, Communication and Knowledge Technologies (ATHENA RC), Greece • University of Edinburgh (UEDIN), UK • University of Sheffield (USFD), UK

For copies of reports, updates on project activities, and other CRACKER-related information, contact:

DFKI GmbH
CRACKER
Dr. Georg Rehm
Alt-Moabit 91c
D-10559 Berlin, Germany

georg.rehm@dfki.de
Phone: +49 (0)30 23895-1833
Fax: +49 (0)30 23895-1810

Copies of reports and other material can also be accessed via http://cracker-project.eu.

Contents

1 Introduction 4

2 SRIA: Main Approach 5

2.1 The Multilingual Digital Single Market 5

2.2 Three Layers 5

3 Time Line and Next Steps 6

4 Document: The Strategic Agenda for the Multilingual Digital Single Market (Version 0.2) 8

5 Presentation: Strategic Agenda for the multilingual Digital Single Market 9

1 Introduction

One objective of CRACKER is to support the EC's plans for establishing a digital single market by means of Language Technologies that enable the *Multilingual* Digital Single Market (MDSM).

To this end, CRACKER and LT_Observatory, a second Coordination and Support Action funded under the Horizon 2020 ICT-17-2014 call are driving the preparation of a Strategic Research and Innovation Agenda (SRIA).

In the past, DFKI had coordinated the preparation of the *META-NET Strategic Research Agenda for Multilingual Europe 2020* that was published in January 2013[1]. The META-NET SRA was elaborated in a long and complex, heavily community-driven preparation and discussion process with multiple feedback loops. The whole process included about 200 contributors (54% from industry; 46% from research) and took more than two years.

The new MDSM SRIA needs to be in a very stable shape in April 2015 when the EC (VP Andrus Ansip and his Project Team on the Digital Single Market) discusses the priorities and main actions towards the Digital Single Market. The details of this discussion process will be presented by VP Andrus Ansip in a press conference at the beginning of May. Shortly before this press conference our Riga Summit 2015 on the Digital Single Market will take place (April 27-29).

As a consequence, the preparation process towards a first stable outline and draft version of the SRIA needed to be very efficient. We decided to take into account all roadmaps and strategy papers that we were aware of (including the META-NET SRA). This is why the SRIA consolidates more than a dozen input documents such as roadmaps, strategic planning documents of visionary presentations of projects and initiatives as well as industry. All input documents and individual contributors are listed in the SRIA.

Several main aspects were discussed with the European Commission as well as with the whole community in the "Workshop on multilingual data value chains in the Digital Single Market" in Brussels on January 16, 2015 (organised by the EC).

Although the title of the Deliverable and the original time line (see Figure 2) suggested to first create a position paper and then turn it into the SRIA, the SRIA editorial team decided to concentrate on the goal of drafting the SRIA right away due to the timing constraints dictated largely by VP Andrus Ansip's plans and communications concerning the time line of his DSM Strategy.

This Deliverable sketches the main approach and time line of the SRIA and includes the SRIA itself (Version 0.2) as well as a presentation that provides more details on the SRIA, its development, main aspects, next steps (with regard to V0.2) and the public consultation phase which is planned for April 2015.

[1] http://www.meta-net.eu/sra

2 SRIA: Main Approach

2.1 The Multilingual Digital Single Market

The SRIA is the Language Technology community's response to the fact that the Digital Single Market doesn't exist yet – instead, the market is fragmented into more than 20 different language communities and sub-markets.

Language Technology (LT) is presented as a critical enabler for the *multilingual* Digital Single Market (MDSM).

Unlike the META-NET SRA, however, the SRIA is not centred around LT, but driven by solutions towards enabling and realising the MDSM, operationalised in a layered approach (see below).

2.2 Three Layers

The setup of the large and ambitious strategic programme towards the MDSM consists of three different layers (see Figure 2).

On the top layer we have a set of focused **Technology Solutions for Businesses, Public Services and Societal Challenges**.

These innovative application scenarios and solutions are, in turn, supported, enabled, and driven by the middle layer which consists of a small group of **Services, Infrastructures and Platforms** that provide, through standardised interfaces, data exchange formats and component technologies, different services for the translation, analysis, production, generation, enrichment and synthesis of written and spoken language.

The bottom layer connects the infrastructures to four innovative **Research Themes**. These research themes provide concrete scientific results, approaches, technologies, modules, components, algorithms etc. that can then be used to enable the second and, ultimately, the top layer.

One additional theme is concerned with **core resources and technologies for language production and analysis**.

This theme touches upon basic technologies for the specific languages to be supported through our programme: in order to equip every language with a set of core resources and technologies, we suggest, among others, intensifying knowledge and technology transfer between larger research centres and groups working on technologies for those languages that are in danger of digital extinction.

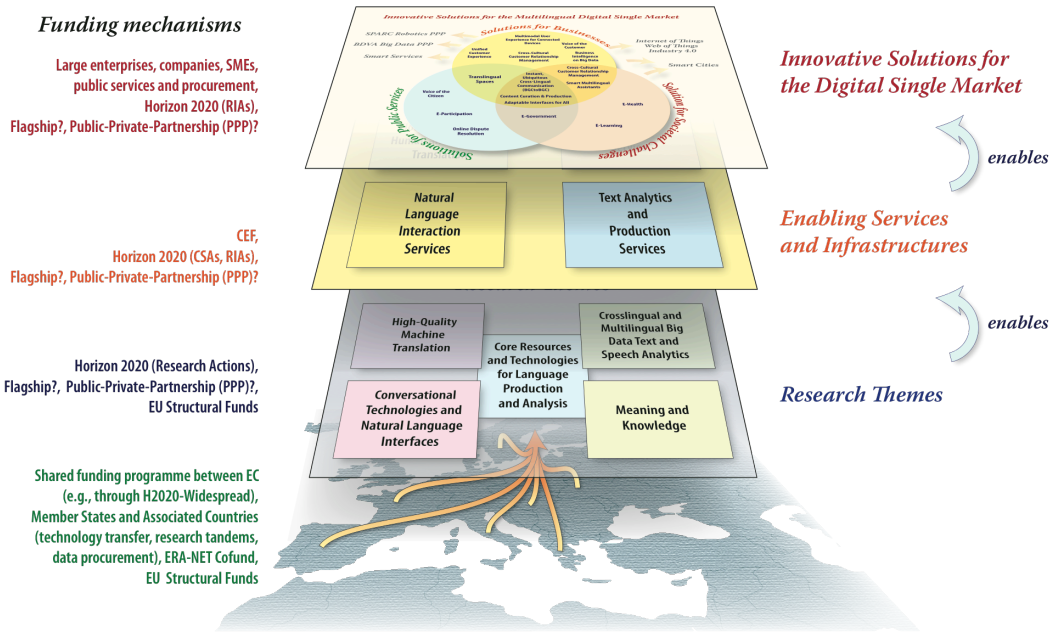


Figure 1: The Layered Approach of the Strategic Programme

3 Time Line and Next Steps

CRACKER started in January 2015. The original time line for the SRIA that was presented at the "Workshop on multilingual data value chains in the Digital Single Market" organized by the EC on 16 January 2015 in Brussels can be seen in Figure 2. A more up to date time line is contained in the presentation that is attached to this deliverable.

Figure 2: Selected inputs and the original time line

The goal of the workshop in Brussels was to collect input and feedback from the participants. In fact, the input and feedback was very valuable and numerous. Together with input from all existing strategy papers (META-NET SRA, LIDER Roadmap, input from ROCKIT, MLi etc.), this made it possible for CRACKER to create a substantial initial draft of the SRIA right away instead of first drafting only a position paper.

An initial SRIA draft, Version 0.1 (V0.1), was circulated to selected collaborators who had volunteered to contribute. Several project consortia (LIDER, MULTISENSOR) provided input and comments at very short notice that have been included in V0.2 (included in this deliverable) that was then circulated to a wide list of recipients for feedback. LT_Observatory also provided feedback and an alternative idea for a section of chapter 2 and the outline of chapter 4. The next steps are to consolidate the input from LT_Observatory with the current concept of the layered strategic programme, to organise a public consultation phase, to present the SRIA at the Riga Summit, to have a second public consultation phase and, afterwards, to finalise the SRIA. However, as the EC's DSM strategy is still work in progress, the currently foreseen timeline towards the finalisation of the SRIA document may change.

As already mentioned, a more current and also more detailed time line is contained in the presentation that is attached to this deliverable.

4 Document: The Strategic Agenda for the Multilingual Digital Single Market (Version 0.2)

Strategic Agenda for the Multilingual Digital Single Market

Solutions for Overcoming Language Barriers for a truly integrated European Online Market

SRIA Editorial Team

Version 0.2 – March 26, 2015

– internal use only –

This draft document is not yet meant to be circulated widely in the community.

Next Steps

- Harmonisation of alternative content suggestions.
- Discussion of the current document with the whole community – public consultation phase. Goal: come to an agreement, as quickly as possible, with regard to the overall setup of the strategic funding programme (three layers, topics, solutions, names and titles etc.). It is mission-critical to get the whole community's support and buy-in.
- Polish, streamline and shorten the sections that are too verbose.
- Complete the missing or partially filled sections.
- Define first priorities (topics, services) and steps including timing, i.e., a very first roadmap.

Notes on Version 0.2

- Outline and overall setup: several changes to the titles and names of solutions. Chapter 2 was modified accordingly (including updated visuals).
- New content in Chapters 3, 4, 5 and 6. (Note: some of the newly included content pieces are indicative only and based on existing content that needs polishing.)
- Included input from Dave Lewis (LIDER, FALCON), especially for Sections 4.2 and 5.4.
- Included a few comments from Stefanos Vrochidis (MULTISENSOR) in Section 5.2.
- Included alternative proposal to the "infrastructures and services" layer (LT_Observatory), see the box in Chapter 2.
- Included edits from Philippe Wacker (LT_Observatory) in Chapter 2.
- Technical glitch: the footnote numbering is broken and needs to be fixed.

Notes on Version 0.1

- This document is an initial, indicative draft.
- Outline reworked substantially.
- Strategic funding programme: prepared overall concept, setup and key visuals.
- Main purpose of Chapter 2 is, for now, to explain the rationale and overall setup, especially with regard to all parties involved in the further preparation of this document.
- Chapter 2 can also be used as the basis of a document to be presented in Riga.
- We took into account all strategy papers, roadmaps and presentations that we had access to (listed in the appendix). Based on these input documents we extracted, structured, generalised and labeled the solutions listed in Chapter 3.
- We've concentrated on the following pieces of content: outline, Chapter 2, overall setup and approach of the programme, solutions (collection and structure), intuitive visuals.
- This document contains an indicative and representative excerpt of the content pieces that we currently have. In Chapters 3 and 4, indicative example content is included (not yet final). We have about 20 additional pages with notes and text pieces.
- Most text is new; some pieces have been taken over from other documents.
- Next steps: agree on structure, scope, setup etc.; identify additional structural pieces for the outline (or for the outlines of later versions of the SRIA); prepare more content until Riga.

Table of Contents

1 Executive Summary 5
2 The Digital Single Market is a Multilingual Challenge 5
2.1 The Digital Single Market and the European Data Economy 8
2.2 The Economic Power of Language Technology and the Language Industry 10
2.3 A Strategic Programme for the Multilingual Digital Single Market 10
2.4 EC and Language Technology – Past and Present 15
2.5 Summary and Conclusions 16
3 Solutions responding to Europe's multilingual Challenges 18
3.1 Technology Solutions for Businesses 19
3.1.1 Unified Customer Experience and Cross-Cultural CRM 19
3.1.2 Voice of the Customer 19
3.1.3 Business Intelligence on Big Data 20
3.1.4 Content Curation and Content Production 21
3.1.5 Multimodal User Experience for Connected Devices 21
3.1.6 Smart Multilingual Assistants 22
3.1.7 Translingual Spaces 23
3.1.8 Ubiquitous Cross-Lingual Communication (BGCtoBGC) 24
3.2 Technology Solutions for Public Services 24
3.2.1 Voice of the Citizen – Social Intelligence on Big Data 24
3.2.2 E-Participation 25
3.2.3 E-Government 26
3.2.4 Online Dispute Resolution 26
3.3 Technology Solutions for Societal Challenges 27
3.3.1 Adaptable Interfaces for All 28
3.3.2 E-Health 28
3.3.3 E-Learning 28
4 Enabling Platforms, Infrastructures and Services 28
4.1 Translingual Trusted Cloud Platform for Human and Machine Translation 30
4.1.1 Implementation 30
4.2 Multilingual Meaning and Knowledge Infrastructure 32
4.2.1 Implementation 34
4.3 Natural Language Interaction Services 34
4.4 Text Analytics and Production Services 34
5 Research Themes 35
5.1 Research Theme 1: HQ Machine Translation and Human Translation 36
5.1.1 Novel Research Approaches and Targeted Breakthroughs 36

5.1.2 Solution and Realisation 37
5.2 Research Theme 2: Crosslingual and Multilingual Big Data Text and Speech Analytics ... 39
5.2.1 Novel Research Approaches and Targeted Breakthroughs 40
5.2.2 Solution and Realisation 42
5.3 Research Theme 3: Conversational Technologies and Natural Language Interfaces 43
5.4 Research Theme 4: Meaning and Knowledge 43
5.4.1 Novel Research Approaches and Targeted Breakthroughs 44
5.4.2 Solution and Realisation 45
5.5 Core Resources and Technologies for Language Production and Analysis 46
6 Horizontal Framework Aspects 47
6.1 Copyright and IPR 47
6.2 Open Source 48
6.3 Language Policy 49
6.4 Standards and Interoperability 50
6.5 Skills 50
7 Organisation of Research 50
8 Conclusions – Summary – Next Steps 50
A. References 52
B. Input Documents 52
C. Detailed Roadmaps 52
D. Digital Language Extinction in Europe 52
E. Key Contributors 53
F. Milestones and History 53

1 Executive Summary

2 The Digital Single Market is a Multilingual Challenge

The Digital Single Market (DSM) holds tremendous potential to transform the European economy and make it more globally competitive. However, one digital "European market" as such does not yet exist: it is instead 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 and it is no surprise that most of the pioneering growth in eCommerce has happened in the US, where regulatory barriers are lower and one language can address the vast majority of the market. Europe needs to face and open up these invisible borders created by one of the most treasured pieces of our cultural heritage: our different languages. All of the languages spoken in Europe are also needed in the Digital Single Market: online shops, information pages, public services, encyclopedias, university pages, company websites, user-generated content, online videos, podcasts, radio stations, and other multimedia content make use of the official, regional, and unofficial minority languages spoken in Europe.

The European Commission predicts that the transition to the integrated DSM will deliver up to €250 billion in economic growth by 2020. However, while eliminating mobile roaming charges, improving telecom, copyright and data protection legislation, and making cross-border payments easier are all important and necessary preconditions for the DSM, they are not sufficient to accomplish the goal. If customers are hampered by language, online commerce will remain confined to fragmented markets, defined 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 more than 20% of the DSM (German comes closest, as the native language of 19% the EU's population). Addressing 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 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, or languages of important trade partners.

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

The Digital Single Market today would account for approximately 25% of global economic potential. However, if Europe can remove the language barriers that hamper intra-European trading, it would also remove barriers to international trade that keep EU-based SMEs from achieving their full

[1] https://www.commonsenseadvisory.com/Default.aspx?Contenttype=ArticleDet&tabID=64&moduleId=392&Aid=21500

[2] EC (2015): http://europa.eu/rapid/press-release_IP-15-4475_en.htm, based on Digital Economy and Society Index

[3] Annual Report on European SMEs 2013/13: A Partial and Fragile Recovery, http://www.eubusiness.com/topics/sme/report-2014/

economic potential by penetrating markets in other continents. 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 in 2013.[6] Most of this increase comes from English, Spanish, and French, but other languages also make significant contributions to world-wide market access. The global potential for European businesses exceeds the internal opportunities from the DSM by orders of magnitude.

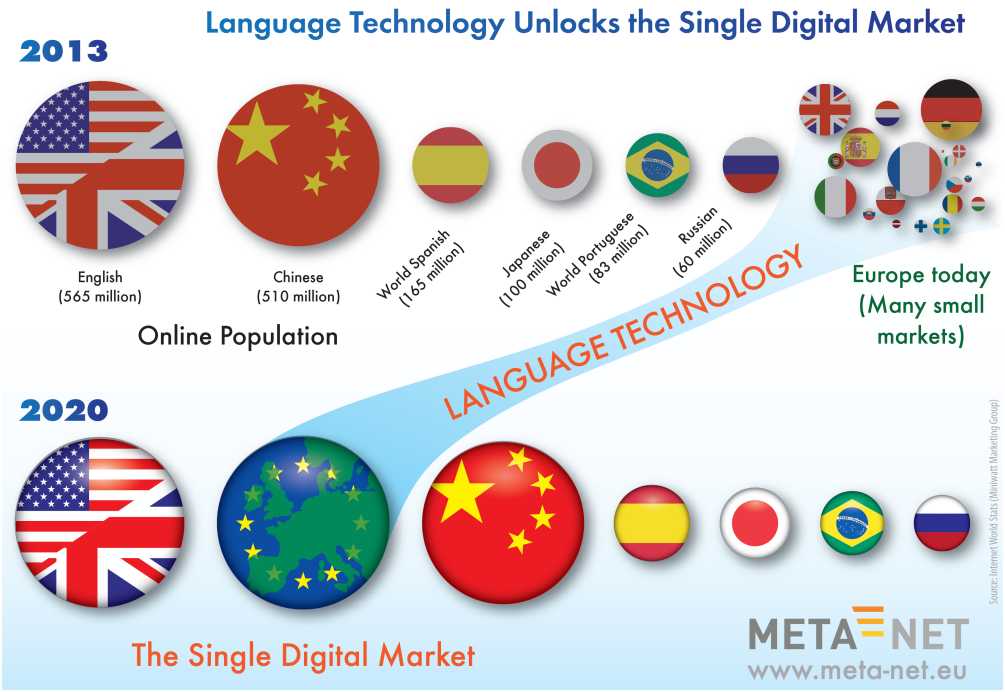
The borders between our beloved languages are invisible barriers at least as strong in their separating power as any remaining regulatory boundaries. They create multiple fragmented and isolated digital markets in which no bridges are provided to other languages, thereby hampering the free flow of products, commerce, communication, ideas, help, and thought. Language barriers of this type in the online world can only be overcome completely by (1) significantly improving one's own skills in non-native languages, (2) making use of others' language skills, (3) or through using 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 content and purposes, specialised human language services, increasingly assisted by language technology, will continue to play a major role, e.g., for translating documents for a fee, creating subtitles for videos, or localising websites into 20+ other languages. However, relying on human services would exclude most SMEs from the DSM because of the high costs. It would create a market that can only be addressed and successfully penetrated by large, consolidated enterprises, which is why cost-effective methods must be found to support market access for SMEs and European citizens.

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 machine translation services offered by large Internet companies are useful for giving users the gist of web content. However, 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 well 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.

The connected and truly integrated Digital Single Market can only exist once all language barriers have been overcome, once 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 the goal of breaking down language barriers. Unfortunately, establishing such a technological infrastructure requires an immense collective push that involves designing and implementing infrastructures, accelerating innovation, basic and applied research as well as technology transfer. While 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, facing the danger of digital extinction (for example, Maltese, Lithuanian and Latvian), even though it must be noted that support for these languages with smaller numbers of speakers is slowly increasing (more details can be found in the Appendix).[8]

[6] Figures from *The 116 Most Economically Active Languages Online*, 2013, Common Sense Advisory.

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



Today's IT systems only start having access to the meaning, purpose and sentiment behind our trillions of written and spoken words. Language makes up a very large part of our big data treasure. Today's computers cannot understand texts and questions well enough to provide translations, summaries or reliable answers in all languages, but in less than ten years such services will be offered for many. Technological mastery of human language will enable a multitude of innovative IT products and services in industry, commerce, government and administration, private and public services, education, health care, entertainment, tourism and many other sectors.

Language technology is 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 – this is why language technology is not an optional but a mandatory component! It is the key enabler and solution to boosting growth in Europe and strengthening our competitiveness in this technology sector that has become so incredibly critical for Europe's future, considering the significance given to the DSM by the European Union.

Europe is the most appropriate place for accomplishing the needed breakthroughs in technology evolution through fundamental and applied research but even more so in technology evolution and profitable innovation. Our continent has half a billion citizens who speak one of over 60 European and many non-European languages as their mother tongue. Europe has more than 2,500 small and medium-sized companies in language, knowledge and interface technologies, and more than 5,000 companies providing language services that can be improved and extended by technology. In addition, it has a long-standing research, development and innovation tradition with over 800 centres performing scientific and technological research on all European and many non-European languages.

Our different European countries and language communities constitute a set of individual, unconnected, fragmented, isolated markets. A truly integrated Digital Single Market that spans our whole 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 vinery, presented in Portuguese, in the first place – due to language barriers and the resulting language-blocking. Furthermore, negotiating and

completing a deal would require a translator.

Current research on the online use of languages demonstrates that the pressure finally to overcome language barriers is increasing. The hitherto dominant languages are retreating as a share of online content and "long-tail" languages are rising.[10] In line with the constant rise of online content, the absolute numbers are rising for all languages but much more significantly so for less common languages. One example in Europe: Basque, Galician, and Catalan 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 an increasing public demand for content in regional or local languages due to the increasing availability of broadband as well as high speed mobile connectivity and the increasing numbers of online users and online services. Europe's citizens are no longer satisfied using a few major languages only. As a consequence, businesses that cannot provide local-language content will be global losers. Furthermore, the numbers indicate that market saturation for dominant languages has been achieved and that any additional growth is coming from outside the established markets historically served by a smaller set of languages.[11] If we extrapolate the trends reported by the think tank Common Sense Advisory, it took 37 languages to reach 98% of the world online population in 2009, and 48 in in 2012. The predicted number in 2015 would be 62 languages. More and more citizens are connected and, as a consequence, more and more citizens use and demand to use their own native languages in any online activities. However, they are excluded from participating in many online activities due to the fact that language barriers constitute market barriers – especially so with regard to the Digital Single Market. True engagement with consumers across language barriers is also deeply entwined with the technical, cultural and individual awareness, preferences and requirements of the user. The power of personalising any cross-linguistic exchange to the individual user means we should not merely bridge the language barrier but provide the compelling tailored user experiences that are key to a vibrant and competitive DSM.

The impact a truly connected DSM could have is not just felt in terms of sales. Technological integration is not helpful if the content contained in systems is not understandable. 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 in another region, e.g., if a traveler 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.

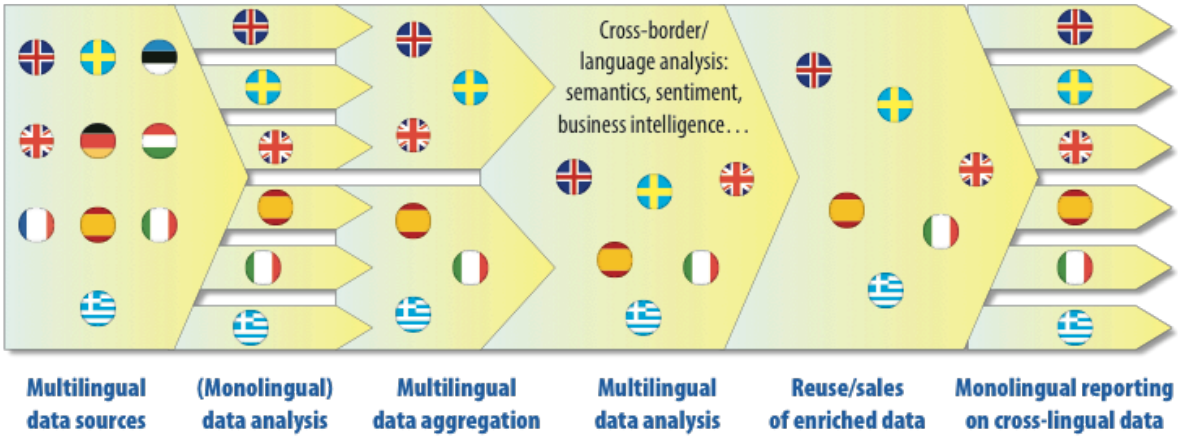
2.1 The Digital Single Market and the European Data Economy

The "language" component is not only a necessary ingredient of the Digital Single Market, it is also a mandatory enabler for the future European Data Economy.

It has been said for a number of years now that data is the oil of the 21st century. Data linking and content analytics are key technologies to refine this oil so that it can drive the engine of many innovative applications through data homogenisation, semantic analysis and repurposing. Here, it is important to note that the large data sets of our Big Data age are never just numerical data – they always come with natural language components such as, for example, column heads in database tables, free text in table cells, metadata annotations, descriptions, documentation, summaries, links to specific documents etc. In other words, not only is the new Data Economy an integral part of the Digital Single Market, to enable an actual Pan-European Data Economy, we need to foresee mechanisms so that data sets and data value chains can flow freely across language boundaries.

[10] 2013, Common Sense Advisory (The Rise of Long-Tail Languages).

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



In addition to the multilingual challenge there are several other technological aspects connected to the Data Economy and Data Analytics. This concerns the sheer volume of data generated. For example, only one hour of customer transaction data at Wal-Mart, corresponding to 2.5 petabytes of data, provides 167 times the amount of data housed for example by the Library of Congress (BilbaoOsorio et al. 2014). The growth rate keeps rising: 90% of the data available today has been generated in the past two years only (SINTEF, 2014). The International Data Corporation 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 (Gantz and Reinsel 2012). The Internet of Things and Web of Things will add even more data and, especially, additional types of data: Cisco estimates that currently less than 1% of physical objects are connected to computer networks. This number will change radically to up to 50 billion devices connected to the Internet by 2020, corresponding to between 6 and 7 devices per person on the planet (Cisco 2013). These examples show that Europe needs a scalable technological infrastructure for handling its big data sets – including robust and precise multilingual text analytics technologies that are able to perform at web-scale level and, even more crucial, at Internet of Things level.

Big data analytics will not just be "slightly better" if we include language technology – it simply will not happen! We cannot download 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 but also to any type of unstructured 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 data value chain. To this end, we suggest to engage in a close collaboration with the Big Data Value Association (i.e., the Big Data cPPP) to ensure the multilingual big data value chain reflects the subtleties and variety of language in the use of vocabulary, register, idioms and irony that is distinct to individuals, communities and domains.

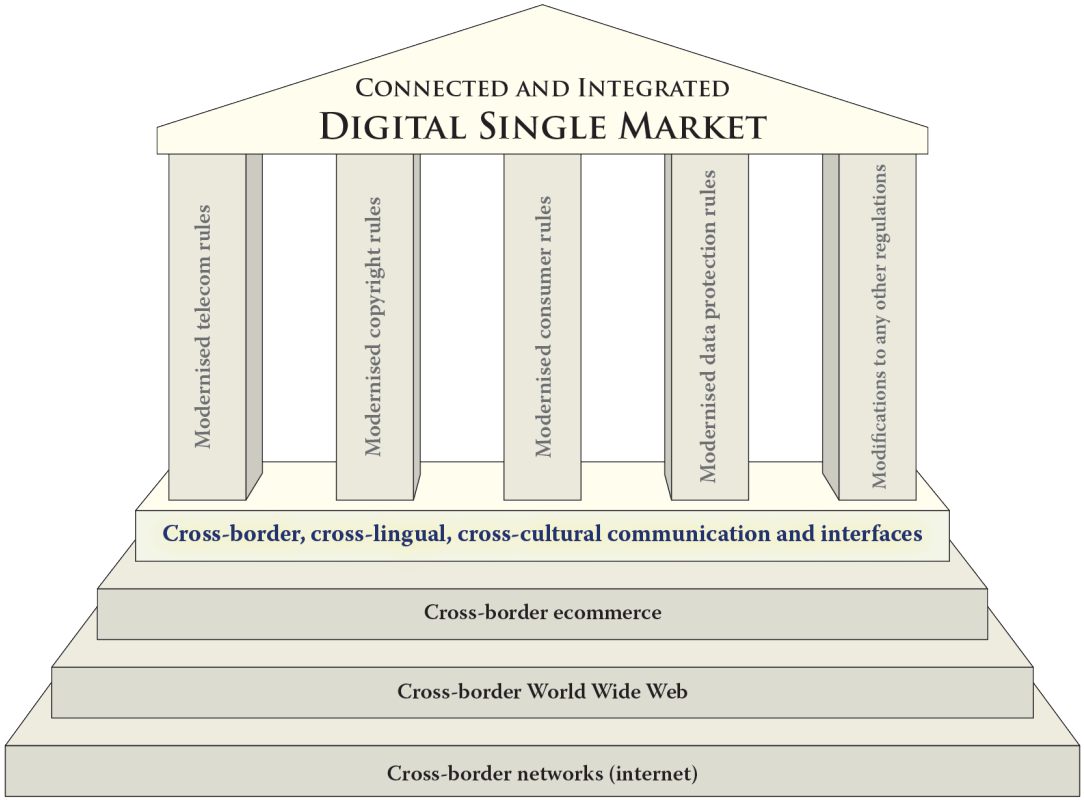
2.2 The Economic Power of Language Technology and the Language Industry

In addition to being a key enabling technology for the multilingual Digital Single Market, the field of 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 speech technology market is even bigger, it 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. Already today, Google Translate translates the same volume per day that all human translators on the planet translate in one year.

FIXME: Include and reference LT Innovate's numbers (more current?).

2.3 A Strategic Programme for the Multilingual Digital Single Market

The integration of the connected Digital Single Market, by definition, must address our different languages. *The Digital Single Market is a multilingual challenge!* Our treasured multilingualism, one of the cultural cornerstones of Europe and one of the main assets of what it means to be and to feel European, is also one of the main obstacles of a truly connected Digital Single Market.



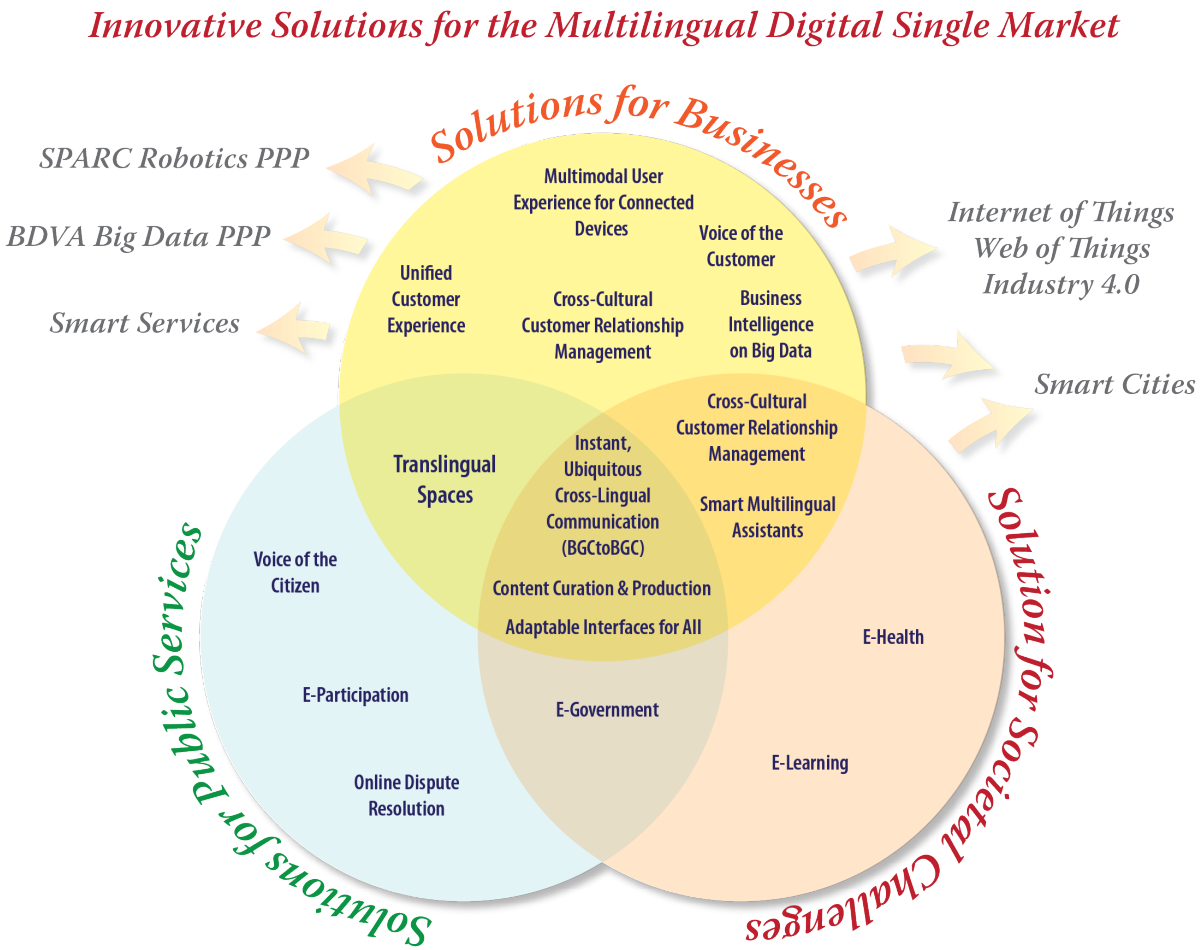
Our goal is to provide the technological facilities for a truly connected and integrated multilingual Digital Single Market through monolingual, crosslingual and multilingual technology support for all languages spoken by a significant population in Europe.

In order to address this challenging goal, we propose a large and strategic programme with a setup that consists of three layers and is firmly grounded in Europe's language communities:

- Layer 1: Innovative Solutions for the multilingual Digital Single Market
- Layer 2: Enabling Services and Infrastructures
- Layer 3: Research Themes

On the **Solutions Layer** (Layer 1) we suggest to focus upon technology solutions for businesses, public services and societal challenges to demonstrate and to make use of novel technologies in solutions with high economic and societal impact and creating numerous new business opportunities for European companies geared towards the multilingual Digital Single Market. While we only briefly list the different solutions here, they are further elaborated upon in Chapter 3.

- **Solutions for Businesses:** Unified Customer Experience; Cross-Cultural Customer Relationship Management; Voice of the Customer; Business Intelligence on Big Data; Content Curation and Production; Multimodal User Experience for Connected Devices; Smart Multilingual Assistants; Translingual Spaces; Instant, Ubiquitous Cross-lingual Communication (Businesses/Governments/Customers/Citizens to Businesses/Governments/Customers/Citizens).
- **Solutions for Public Services:** Voice of the Citizen – Social Intelligence on Big Data; E-Participation; E-Government; Online Dispute Resolution.
- **Solutions for Societal Challenges:** Adaptable Interfaces for All; E-Health; E-Learning; Elder Care; Cultural and Heritage Preservation; Environmental Management & Preservation.



On the **Enabling Services and Infrastructures Layer** (Layer 2) we suggest to establish a small group of infrastructures, platforms and services to connect the innovative technology solutions (see above) with foundational and applied research activities (see below).

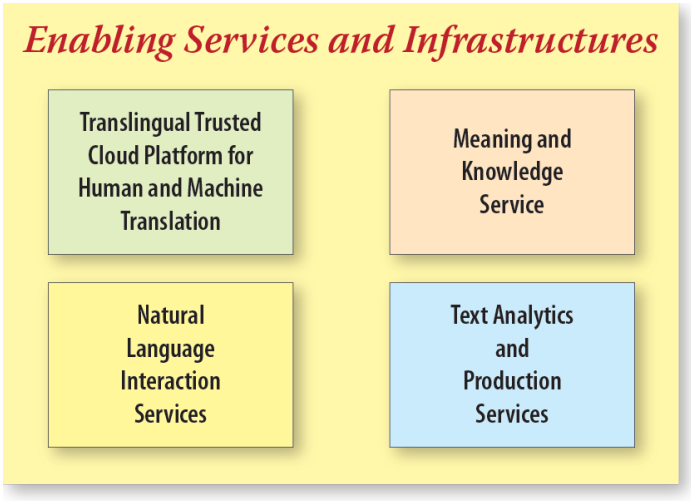
First, a **Translingual Trusted Cloud Platform for Human and Machine Translation** is needed as one of the core technological services of our suggested programme. This platform must be designed from the outset with special emphasis on high-quality output, trust, data security, reliability, privacy, data protection and confidentiality.

Second, a **Multilingual Meaning and Knowledge Service** needs to be realised. This service provides seamless and ubiquitous access to a multilingual knowledge base that integrates information about products, companies, places, terms, words, and a plethora of other concepts that are of vital importance for all monolingual, crosslingual and multilingual language technology components and data value chains. Designing and implementing such a knowledge service is a challenge but it can become a reality through the combination of existing repositories such as Wikipedia, Wikidata, DBPedia, Linked Open Data sets, WordNet and many other language and data resources.

Third, we need at least one (or maybe several) rather generic service platform for **Text Analytics and Production Services**. Some of the language technology methods and components assembled in this platform (tokenisation, POS-tagging, parsing, spell-checking etc.) are so fundamental and generic that they can and should be provided under one umbrella. This platform also includes more complex and sophisticated methods for text analytics, report generation, text classification, sentiment analysis and opinion mining.

Fourth, a platform for **Natural Language Interaction Services** is needed to assemble all sorts of services for the analysis and synthesis of spoken language such as, among others, automatic speech recognition and text-to-speech methods. Just like all the other infrastructures and platforms, the platforms need to be interconnected. The speech platform, for example, especially needs to contain bridges to the translation cloud and multilingual meaning and knowledge services.

All platforms need to have a 24/7 availability and provide web-scale performance and connectivity in order to perform their main purpose: to support research and innovation by testing and showcasing results as well as providing an environment for hybrid research, i.e., integrating research and operational services. These infrastructures will allow providers from research and industry to offer services, resources and component technologies.



LT_Observatory provided an alternative proposal concerning the "Enabling Services and Infrastructures" layer (Layer 2). We include this alternative proposal in the following four paragraphs so that both proposals can be discussed.

On the **Enabling Services and Infrastructures Layer** we suggest to establish a small group of infrastructures, platforms and services to connect the innovative technology solutions (see above) with foundational and applied research activities (see below).

First, Europe needs a basic infrastructure for natural language processing, the **European Language Cloud**. All language processing applications (search, mining, writing, speech, translation, etc.) depend on such basic infrastructure. These are tedious to develop and to maintain, and expensive, since they are required for every single language. The European Language Cloud (ELC) is a public infrastructure which provides the basic functionality required to process unstructured content. Through an API it provides basic language technology services such as tokenization, stemming, part of speech tagging, named entity detection, Identification of measurements, currencies, formulas, etc. for all languages in the same base quality under the same favourable terms.

Second, a **Multilingual Meaning and Knowledge Service** needs to be realised. This service provides seamless and ubiquitous access to multilingual knowledge bases that integrates information about products, companies, places, terms, words, and a plethora of other concepts that are of vital importance for all monolingual, cross-lingual and multilingual language technology components and data value chains. Designing and implementing a general knowledge service is a research challenge but it can become a reality through the combination of existing repositories such as Wikipedia, Wikidata, DBPedia, Linked Open Data sets, WordNet and many other language and data resources. Important for industry and eGov will be sector-specific multilingual knowledge systems which are key assets for serving a global customer base or achieving semantic interoperability.

Third, we need a series of **European Language Technology Application Platforms** for verticals as generic but sector-specific infrastructures. These platforms will provide key industries with a range of language tools and resources that are specifically tailored to the knowledge, linguistic and business process needs of the industry in question. They should be built in coordination with major companies operating in vertical sectors that see an advantage to sharing certain resources with their competitors so as to avoid "reinventing the wheel" and becoming more competitive individually and as a sector. Language technology suppliers who provide the services on these platforms will be able to draw on research and innovation outcomes from other layers to ensure that the technology remains cutting edge. Typical verticals that would be candidates for these services could be automotive, various parts of the healthcare sector, chemicals, legal and financial services, media & publishing, construction.

On the **Research Layer** (Layer 3), which subsumes both basic and applied research, four broad themes drive research and innovation. With regard to these themes it is important to note that all themes are tightly intertwined, making use of one another in different application scenarios, especially so when research results, i.e., technologies, are combined on the solutions layer.

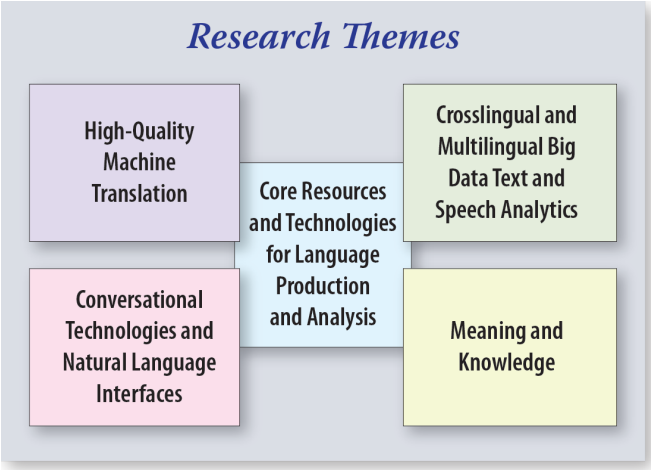
First, the theme **High-Quality Machine Translation including Human Translation** will provide research results, algorithms, approaches, services, and scientific output meant to be directly made use of within the layer of generic and specialised federated cloud services for reliable spoken and written translation among all European and major non-European languages.

Second, under the theme **Crosslingual and Multilingual Big Data Text and Speech Analytics** research will be carried out towards understanding and dialogue within and across communities of citizens, customers, clients and consumers. This includes, among others, research scenarios towards multilingual sentiment analysis, opinion mining, fact mining, rumour detection, information and relation extraction as well as components that construct semantics for linguistic analyses through a connection to the "Meaning and Knowledge" theme – always taking into account the multitude of established and emerging online text types and genres.

Third, within the theme **Conversational Technologies, Dialogue Systems, Natural Language Interfaces** we suggest to intensify research on speech interfaces and interactive assistants – for all European languages. Especially with regard to the Internet of Things and Web of Things as well as trends such as Wearables and Advanced Manufacturing, a very high demand for natural language interfaces can be predicted for the near future. These interfaces also include socially-aware interactive and pervasive assistants that learn and adapt and that provide proactive and interactive support tailored to the respective user's context.

Fourth, the theme **Meaning and Knowledge** provides an umbrella for aligning and harmonising all research activities around monolingual, crosslingual and multilingual resources, data sets, repositories, and knowledge bases that are needed as background knowledge for all advanced language processing components – from machine translation to text analytics to speech interfaces. For example, this theme needs to 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. All existing resources need to be consolidated, made interoperable, aligned and enriched with multilingual information. Additionally, research needs to work on novel 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 soon. 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 the knowledge is contextual, and the user must be taken into account in a manner that is first and foremost privacy preserving, retains user control and yield transparent protection of user data.

The final building block of this layer is concerned with providing **Core Technologies and Resources for Europe's Languages**. We propose to realise a system of shared, collectively maintained, interoperable tools and resources that will ensure that our languages will be sufficiently supported and represented in future generations of IT solutions. This system of shared tools and resources is a crucial prerequisite for the multilingual Digital Single Market because it connects our programme to the different languages. Many of these core technologies and resources will be made available as services through one or more of the European Service Platforms for Language Technologies.



This three-layer approach will provide European research, development, and innovation in the field of innovative language technologies and also multiple different industries with the ability to compete with other markets and subsequently achieve multiple benefits for European society and citizens as well as an array of opportunities for our economy and future growth.

2.4 EC and Language Technology – Past and Present

In the late 1970s the EU realised the profound 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 a few years ago; in an internal reorganisation this department was integrated into a new unit called “Data Value Chain”, part of Directorate G, “Media & Data”, in the EC Directorate General for Communications Networks, Content and Technology (DG Connect).

In the past ca. ten years, the EU has been supporting projects such as, for example, EuroMatrix and EuroMatrixPlus (2006-2008, 2009-2012) as well as iTranslate4 (2010-2012), which use basic and applied research along with industrial collaborations to generate resources for establishing high-quality machine translation solutions for all European languages such as, among others, the Moses system. More recently, the large-scale initiative META-NET (supported, in its first phase, through 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 750 members. META-NET has prepared studies such as the more than 30 volumes of the META-NET White Papers Series, and the META-NET Strategic Research Agenda for Multilingual Europe. Furthermore, it devised the open resource exchange facility META-SHARE which makes available more than 2,500 language resources. 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[16] and an Innovation Agenda[17].

At the beginning of 2015 new projects have been launched, funded through the call Horizon 2020-ICT 17, “Cracking the Language Barrier”. 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 several industrial and commercial use cases.

In parallel to the research and innovation-oriented activities funded through FP7 and Horizon 2020, the European Commission is further advancing the Connecting Europe Facility programme (CEF). Part of CEF Digital 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

[16] Unleashing the Promise of the Language Technology Industry for a Language-neutral Digital Single Market, June 2014
[17] LT2013: Status and Potential of the European Language Technology Markets, April 2013

all of CEF's Digital Service Infrastructures multilingual.[18] This automated translation service, CEF AT, builds on an existing machine translation system, MT@EC, developed at the EC (DG Translate), based on the existing Moses system, under the Interoperability Solutions for European Public Administrations (ISA) programme with support of FP7 and CIP. 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. Availability of MT@EC is currently restricted to staff members of the EC and the EP. During 2015 a closer collaboration between CEF AT and the European language technology community will be established, especially with regard to the systematic and coordinated collection and exploitation of language resources in all countries.

Looking beyond the EC, research by TAUS[19] has shown that European research funding that fostered the development of the open source machine translation 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. The long-term trend to increasing language support and increasing customer engagement via language technology is clear. According to the report, there are already 22 Moses-based MT companies operative with an estimated market share of about $45 million or about 20% of the entire MT solutions market.

2.5 Summary and Conclusions

For the large and ambitious strategic funding programme we recommend a setup that consists of three different layers: on the top layer we have a set of focused **Technology Solutions for Businesses, Public Services and Societal Challenges**. These innovation application scenarios and solutions are, in turn, supported, enabled, and driven by the middle layer which consists of a small group of **Services, Infrastructures and Platforms** that provide, through standardised interfaces, data exchange formats and component technologies, different services for the translation, analysis, production, generation, enrichment and synthesis of written and spoken language. The bottom layer connects the infrastructures to four innovative **Research Themes**. These research themes provide concrete scientific results, approaches, technologies, modules, components, algorithms etc. that can then be used to enable the second and, ultimately, the top layer. One additional theme is concerned with **core resources and technologies for language production and analysis**. This theme touches upon basic technologies for the specific languages to be supported through our programme: in order to equip every language with a set of core resources and technologies, we suggest, among others, intensifying knowledge and technology transfer between larger research centres and groups working on technologies for those languages that are in danger of digital extinction.

[18] https://joinup.ec.europa.eu/community/cef/og_page/catalogue-building-blocks#AT

[19] *Moses MT Market Report*, TAUS, 2015 (https://www.taus.net/think-tank/reports/translate-reports/moses-mt-market-report)

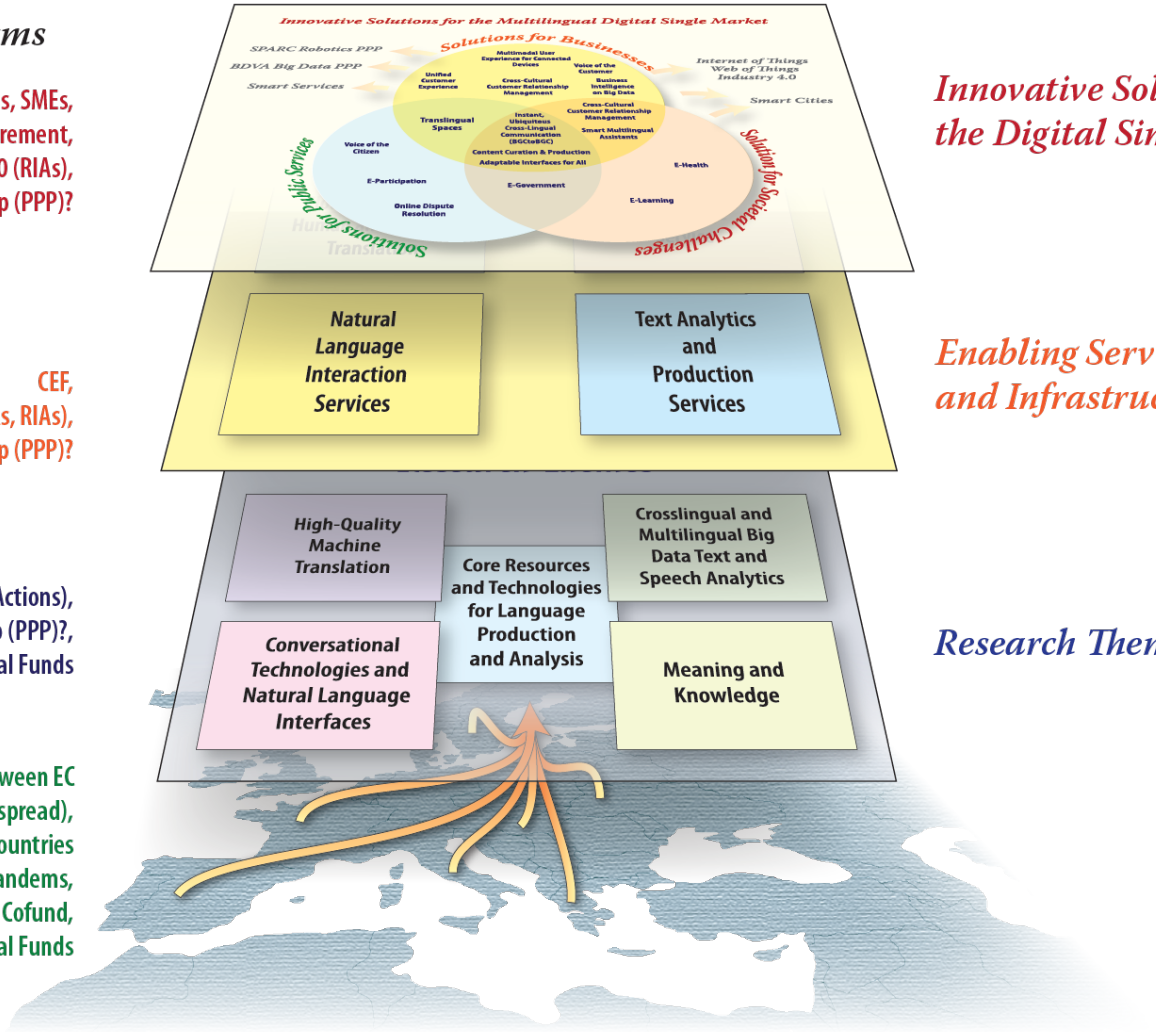
Funding mechanisms

Large enterprises, companies, SMEs, public services and procurement, Horizon 2020 (RIAs), Flagship?, Public-Private-Partnership (PPP)?

CEF, Horizon 2020 (CSAs, RIAs), Flagship?, Public-Private-Partnership (PPP)?

Horizon 2020 (Research Actions), Flagship?, Public-Private-Partnership (PPP)?, EU Structural Funds

Shared funding programme between EC (e.g., through H2020-Widespread), Member States and Associated Countries (technology transfer, research tandems, data procurement), ERA-NET Cofund, EU Structural Funds



Innovative Solutions for the Digital Single Market

enables

Enabling Services and Infrastructures

enables

Research Themes

This three-layer approach will provide European research, development, and innovation in the field of innovative language technologies and also multiple different industries with the ability to compete with other markets and subsequently achieve multiple benefits for European society and citizens as well as an array of opportunities for our economy and future growth.

An integral component of our strategic plans are the member states and associated countries: it is of utmost importance to set up, under the overall umbrella of the strategic funding programme proposed in this document and its three-layer approach, **a coordinated initiative both on the national (member states, associated countries, regions) and international level (EC/EU), including research centres as well as small, medium and large enterprises who work on or with language technologies**. One instrument for such a coordinated initiative could be setting up a contractual Public-Private-Partnership (PPP). A European Flagship Project could also be a good candidate instrument, especially with regard to significantly boosting the development of innovative and novel LT approaches, algorithms and paradigms as well as supporting the fight against digital language extinction.

Only through close cooperation between all stakeholders, tightly coordinated collaboration and an agreement as well as update of our national and international language policy frameworks can we realise the ambituous plan of researching, designing, developing and rolling out platforms, services and solutions that support all businesses, public services and citizens of Europe, and beyond, and that also enable the multilingual Digital Single Market.

What is missing in Europe is awareness, political determination and political will that would take us to a leading position in this technology area through a concerted funding effort. This major dedicated push needs to include the political determination to modify and to adopt a shared, EU-wide language policy that foresees an important role for language technologies.

As Europeans, we urgently have to ask ourselves a few crucial questions: Can we afford our information, communication and knowledge infrastructure to be highly dependent upon monopolistic services provided by US companies (technological lock-in)? What is Europe's fallback plan in case the language-related services provided by US companies that we rely upon are suddenly switched off or if even more serious access or security issues arise? Are we actively making an effort to compete in the global landscape for research and development in language technology? Can we 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? Can the European cultural background help shape the knowledge society by offering better, more secure, more precise, more innovative and more robust high-quality language technology?

We believe that *Language Technology made in Europe for Europe* will significantly contribute to future European cross-border and cross-language communication, economic growth and social stability while establishing for Europe a worldwide, leading position in technology innovation, securing Europe's future as a world-wide trader and exporter of goods, services and information. Only a large, coordinated push of this magnitude will be able to unlock a truly multilingual Digital Single Market.

3 Solutions responding to Europe's multilingual Challenges

This chapter describes the technology solutions for business (Section 3.1), public services (Section 3.2) and societal challenges (Section 3.3) that will enable a truly integrated multilingual Digital Single Market. The technology solutions described in this chapter – Layer 1 of the three-layer-setup – are enabled by the platforms, infrastructures and services specified in Chapter 4 – Layer 2 of the three-layer-setup. These are, in turn, enabled by the research themes discussed in Chapter 5 – Layer 3 of three-layer-setup.

3.1 Technology Solutions for Businesses

In this section we sketch several technology solutions for businesses, all of which relate to enabling the multilingual Digital Single Market, making it a reality. The "multilinguality component" is an inherent part of all solutions, which is why this specific keyword is not mentioned explicitly in all solution titles described here. The technology solutions relate to aspects such as, for example, a unified customer experience in ecommerce scenarios, market research, business intelligence, content curation and content production. These technology solutions are meant to be developed by commercial solution providers, based, first and foremost on the platforms, infrastructures and services provided by the second level of the three-level-architecture. The primary customers of these solution providers are, again, commercial companies and other organisations that have a need for the respective technologies within the multilingual Digital Single Market.

3.1.1 Unified Customer Experience and Cross-Cultural CRM

- This solution includes crosslingual ecommerce, localisation, internationalisation, translation, processing and generation of written/spoken language, multilingual online marketing etc.
- Includes, like all solutions, "multilingual" by default without mentioning it in the title all the time.
- This solution provides a contextualized, unified digital experience to users, bringing together content, product, customer care, customer relationship (CRM), discussion fora, helpdesks etc. in a unified digital (eco)system across languages.

In today's hyper-connected society, consumers expect to quickly and easily get what they need from your business – anytime, anywhere. This includes access to products and services, but also to information and to easy-to-use, powerful self-services. Today, 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, languages and audiences.

Automation helps bring together content, product and customer relationship management in one ecosystem. The goal is a seamless network of data and knowledge that spans multiple modalities and channels (IVR, mobile, web etc.) and incorporates open and closed datasets in a way that is respectful to intellectual property (IP), data privacy and corresponding licenses. Realising a certain degree of 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. The creation of these resources should be based on standardised ways for representing and linking such information.

In cross-cultural customer relationship management, integrating translation technologies (also supporting niche languages used in micro-domains) into the Customer Engagement Ecosystem 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, contextualized digital experience to all users.

3.1.2 Voice of the Customer

- This technology solution provides comprehensive methods for multilingual market research.
- Extract and interpret the multilingual "voice of the customer" with a high level of accuracy, across languages and modalities, and analysing sentiment at deeper levels beyond mere

polarity, including intention recognition.

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 natural 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. The "Voice of the Customer" technology solution includes the (targeted) analysis of large volumes of such comments and communications created by citizens, customers, patients, employees, consumers and other stakeholder communities. Summarising multiple data streams in real time involves dealing with high-volume, high-velocity data, often of unknown veracity.

Social media analytics builds on improved text analytics methodologies but goes far beyond the analysis. Part of the analysis is directed to the status, opinions and acceptance associated with the individual information units. 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 (beyond mere polarity) is a crucial component of social intelligence.

Textual content analytics will play a crucial role in areas like

- analysing the voice and actions of the customer in the context of CRM,
- brand, product and reputation management,
- technology monitoring and competitive intelligence,
- content management and publishing, and
- search, information access and question answering.

Automatic 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 with a high level of accuracy and across natural languages and modalities, while improving how culture and individual 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.

3.1.3 Business Intelligence on Big Data

- This technology solution includes the crosslingual analysis of big data, multilingual report generation, summarisation, trend detection, question answering, semantic search etc.
- This solution is one of the topics that provide a bridge to the Big Data Value Association cPPP.

The quality, speed and acceptance of individual and collective decisions is the single main factor for the success of enterprises (and, likewise, public services, communities, states and supranational organisations). The growing quantity and complexity of accessible relevant information poses a serious challenge to the efficiency and quality of decision processes. IT provides a wide range of instruments for intelligence applications. Business intelligence, military intelligence or security intelligence applications collect and pre-process decision-relevant information. Analytics applications search data for information and decision support systems evaluate and sort the information and apply problem-specific decision rules.

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 10 billion US$ 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 extraction, document classification, entity extraction, relation extraction, event 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 the big language 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 European 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. We therefore suggest to engage in a close collaboration with the Big Data Value Association cPPP.

3.1.4 Content Curation and Content Production

- This technology solution relates to intelligent authoring support, multilingual and multimodal article, report and text generation, cross-lingual content, content linking, content enrichment and content semantification, automatic subtitling etc.

Collecting, organising 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 processing large amounts of information in a short time. Content curation reduces the overall flow of information and makes it more targeted to the end user's interests, for example for selecting information that is appropriate for corporate blogs or websites or content for brands to post to social media channels. Language Technology solutions can play a crucial role in this process. 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.).

Language technology can also be of help in various content production tasks. Standardised communication, for example email communication in customer support, can be automated by analysing the user input at the meaning level and identifying relevant, semantically similar previous communication. Robot journalism can comb structured data for facts and trends and combine them with contextual information to form sentences, enabling the computer-assisted generation of multilingual articles, reports or product websites, also taking into account other data sources such as, for example, website access analytics. 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 to proactively suggest completions of started sentences or even whole paragraphs. Advanced technologies can check for appropriate style according to genre and purpose and help improve comprehensibility.

3.1.5 Multimodal User Experience for Connected Devices

- This technology solution relates to multilingual speech, text, and gesture interfaces for IoT, WoT, Industrie 4.0, robots, cars, household appliances, consumer products etc.

Virtually all information and knowledge will soon be available in digital form – as a result the

volumes of information about the world are growing exponentially. The consequence is a gigantic distributed digital model of our world that is continuously growing in complexity and fidelity. Through massive networking of this information and the linking of open data, this knowledge repository is getting more useful as a resource for information, planning, and knowledge creation. Among other methods, machine learning enables us to tackle the fusion of this knowledge with signals from different interaction modalities to better understand the user's intent and utterances.

Many everyday objects are already connected to the internet and may even be interconnected with other objects (Internet of Things, Web of Things). Depending on the function, complexity, relevance, and autonomy of these objects, the nature of desired or needed communication can vary widely. Some objects will come with interesting textual information that we would like to query and explore, such as manuals and consumer information. Other objects will provide information on their state and will have their own individual digital memory that can be queried. Objects than can perform actions, such as vehicles and appliances, will accept and carry out (multilingual) voice, gesture or eye-tracking commands. Non-invasive wearable biometric sensors can provide signals about a person's mood or emotional state, offering new affect-focused multimodal interaction with devices. Objects will therefore offer more engaging interaction experience with the person through the combination and optimisation of the multiple modalities available to present digital content and sensed human engagement.

Robots are currently evolving into collaborative, social machines that will eventually provide useful services to humans in numerous work, medical, educational and household contexts. Specialised mobile robots will be deployed for personal services, rescue missions, household chores, and tasks of guarding and surveillance. The effort to design social awareness and the capacity to learn into a robot's computer is delivering returns, as humans and their robots can now team up and share out tasks more intelligently. Physical environments such as offices, homes, hospitals and other sites will, like robots, be able to learn about humans and discover unforeseen needs.

With the evolution of connected devices and robots, enabling communication via natural language commands and dialogues will be the major challenge. Building on top of existing technologies for natural language interaction, including dialogue management and speech technologies, will help create devices that can communicate with us in human language in a user-friendly way. Their wide acceptance will improve productivity, safety and comfort.

3.1.6 Smart Multilingual Assistants

- This solution relates to smart multilingual assistants as they are used in wearables, mobile phones, tablets, connected devices etc.

With the growing number of assisting software in phones, tablets and other connected devices, there is a rising need for socially aware, highly personalised assistants that learn and adapt and that provide proactive and interactive support tailored to specific situations, locations and goals of the user. Voice, gender, language, and mentality of the virtual character need to be adjusted to the user's preferences. The personality and functionality of the interface may also depend on the user type: there may be special interfaces for children, foreigners, and persons with disabilities. Having been trained on the user's behaviour, digital information, and communication space, the assistants can proactively offer valuable unrequested advice.

There will be a competitive landscape of intelligent interfaces to the offered services employing human language and other modes, such as manual and facial gestures, for effective and affective communication. Natural language is by far the best communication medium for interacting with virtual assistants. Integrating multilingual speech technologies will enable the assistants to speak in the language and dialect of the user, but also digest information in other natural and artificial languages and formats, and may even translate or interpret without the user having to request it.

Depending on the needed functions and available information, language coverage will range from simple commands to sophisticated natural dialogues with avatars through speech, tone and

gestures, as illustrated below:

- “Book a table at the Capital Grille after my last meeting ... oh, and let Tom know to meet me there.”
- “Sorry, nothing’s open, but L’Andana is available at 6:30pm”

The next generation of smart assistants will have to achieve a high out-of-the-box accuracy, but will also have to be able to deal with composite meta-tasks with dynamic collaborative interactions, which will require deep natural language understanding and the ability to reason on knowledge. To achieve such smart assistants and conversational agents it is necessary to better model, synthesise and understand social speech signals, e.g., laughter, backchannels, pause insertion and duration, intonation, turn taking, etc. The assistant will translate or interpret without the user even needing to request it.

In the future, many providers of information about products, services, or touristic sites will try to present their information with a specific look and feel. The personality and functionality of the interface may also depend on the user type: there may be special interfaces for children, foreigners, and persons with disabilities. New levels of audio and visual resource management and synchronisation will be needed to handle the variety of body and voice features related to the personality and affect (e.g. emotions, laughter) to offer more human-like assistant technology and to open up new horizons in the generation of creative content (e.g., computer games, movies, music, internet). However, challenges remain for speech recognition to deal with noisy environments, multiple speakers, localising the speaker, as well as understanding conversations and non-verbal signals.

Multilingual assistants are closely related to the Internet of Things. Sensors and power-efficient signal processing are critical for real-world usability. This includes intelligent wake-up functions (‘always listening’), secure biometrics, convenience, context for accurate interpretation, multi-microphone beam-steering, audio-visual recognition, touch and gestures, location, time, movement, vitals (healthcare applications) etc.

3.1.7 Translingual Spaces

- This solution relates to interactive meeting rooms with added services, ambient translation, automatic note taking etc.

As businesses and other organisations attempt to cut down on high-carbon travel agendas, they are turning to virtual meetings online as a cost-effective solution for collaborative events, especially those that are urgent and involve remote locations. Virtual tele-meetings utilising large displays and comfortable technology will be the norm for professional meetings. Tremendous value can be added to such meetings by providing automated interpretation solutions for spoken communication or by automatically transcribing and eventually summarising the content of meetings. Incrementally drafted (searchable) summaries will be used for displaying the state of the discussion, including intermediate results and open issues, and to generate meeting minutes. Brainstorming will be facilitated by semantic lookup and structured display of relevant data, proposals, charts, pictures, and maps from databases and will enable meeting members to make more relevant contributions. Individual realtime translation will simultaneously interpret the contributions of participants, slides, and handwritten text (e.g., on a shared whiteboard) into as many languages as needed. These ‘learning’ services can streamline the entire meeting process, save time, produce an automatic record, and improve teamwork.

Apart from meeting management, the described technologies have the potential to empower and augment human-human communication in general. We envisage applications which help people to be more creative more of the time (especially in group situations), new approaches to social sharing (across languages), design-enabling platforms which enable people to build their own tools, and systems to enable groups (at all scales) to collaborate with shared goals. These applications will facilitate problem solving and provide powerful mechanisms for engagement.

Example use cases include

- Social platforms that encourage people to share and collaborate
- Enhanced meetings (remote and face-to-face) with text and speech translation, just-in-time recommendation, etc.
- Applications for visual, multimodal and audio artistic creativity
- Games (entertainment and serious games)
- Shared (task) understanding and communication, for example in multi-discipline professional teams (e.g., medicine)
- Shared learning, MOOCs, peer-to-peer learning

The underlying technologies will be guided by partial understanding of the contents, i.e., by its semantic association with concepts in semantic models of domains and processes. Reliable dialogue translation for face-to-face conversation and telecommunication will require high-quality and high-speed machine translation that is available for many languages, across multiple subject fields and text types, both spoken and written.

3.1.8 Ubiquitous Cross-Lingual Communication (BGCtoBGC)

- This solution relates to machine translation services, free (for the citizen) or for a fee (specialized HQ services)
- From and to Business, Government, Customer, Citizen (BGC)
- This solution can be conceptualised as the user-facing service side of the Translingual Trusted Cloud infrastructure

Translation services will move 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 will 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 translation, terminology, dictionaries, checking, TMs) and free human translation or post-editing services for special purposes provided by NGO-initiatives, e.g., Translators without Borders, Rosetta Foundation. This technology solution foresees one or more common, easy-to-use access points for citizens, professionals, businesses, and public organisations providing ubiquitous and instant access to information and communication in any language.

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. Such a technological solution will be critical to accelerating the emergence of the multilingual Digital Single Market and other trading platforms.

3.2 Technology Solutions for Public Services

3.2.1 Voice of the Citizen – Social Intelligence on Big Data

- This solution includes large-scale, web-scale sentiment analysis, opinion mining, multilingual report generation, trend analysis; decision support, to increase social reach and approach cross-cultural understanding, to create a “citizen experience” – as a complement to “customer experience” or “user experience”.
- This solution is a complement to the “voice of the customer”.
- Democracy will be enriched by powerful new mechanisms for developing improved collective

solutions and decisions (also see E-Participation).

3.2.2 E-Participation

Today, any collective discussion processes involving large numbers of participants are bound to become non-transparent and incomprehensible, especially as they span the myriad linguistic and cultural boundaries that characterize Europe. Since many collective discussions will involve participants in several countries, e.g., EU member states, cross-lingual participation needs to be supported. By recording, grouping, aggregating and counting opinion statements, pros and cons, supporting evidence, sentiments and new questions and issues, the discussion can be summarised and focused across boundaries to aid engagement across the EU electorate. Decision processes can be structured, monitored, documented and visualised, so that joining, following and benefitting from them becomes much easier. The efficiency and impact of such processes can thus be greatly enhanced. Special support will also be provided for participants not mastering certain group-specific or expert jargons and for participants with disabilities affecting their comprehension.

Meaningful EU-wide citizen engagement in EU issues is not however just restricticted by language barriers. Cultural outlook and national viewpoints still dominate public discourse. Promoting deeper cultural and historical understanding across Europe is therefore key to building a strong sense common history and identity, especially in confronting the influence of previous strife between national, regional and ethnic groupings. The EU has successfully invested promoting public online access to cultural and historical resources through initiatives such as CLARIN and DARIAH, however these resources are still largely contained in linguistic siloes. Language technologies such as machine translation and cross-lingual search can assist, but solutions require extremely careful translation of person, place and event names, e.g., ‘Danzig’ versus ‘Gdansk’, or ambiguous word such as the German word ‘Gewalt’ meaning legitimate power or violence depending on the context. Existing knowledge resources, such as Wikipedia, can help contextualise the significance of words to different national or cultural audiences, but solutions are needed to ensure the language resources used by language technology accurately capture and maintain meta-data on audience sensitivities.

Social Intelligence will support understanding and dialogue within and across communities of citizens and consumers to enable e-participation and more effective processes for preparing, selecting and evaluating collective decisions. The quality, speed and acceptance of individual and collective decisions is the single main factor for the success of social systems such as communities, public services, states and supranational organisations. The growing quantity and complexity of accessible relevant information poses a serious challenge to the efficiency and quality of decision processes. 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 10 billion US$ 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.

Social intelligence builds on improved text analytics methodologies but goes far beyond the analysis. One central goal is the analysis of large volumes of social media, comments, communications, blogs, forum postings etc. of citizens, customers, patients, employees, consumers and other stakeholder communities. Part of the analysis is directed to the status, opinions and acceptance associated with the individual information units. As the formation of collective opinions and attitudes is highly dynamic, new developments need to be detected and trends analysed. Emotions play an important part in individual actions such as voting, buying, supporting, donating and in collective opinion formation, the analysis of sentiment is a crucial component of social intelligence.

We envision the emergence of one or more public discussion and opinion formation platforms for Europe-wide deliberations on pressing issues such as energy policies, the financial system, migration, natural disasters, which also supports the proactive engagement of less active parts of

the population. Addressing politicians, health providers, manufacturers, the cultural sector and citizens, it will provide visualisations of social intelligence-related data and processes for decision support. Solutions will be based on the detection and prediction of events and trends from content and social media networks and on mining e-participation content for recommendations and summarisation. Building the required technologies will require the usage of high-throughput, web-scale content analysis techniques that can process and extract knowledge from multiple different sources, ranging from unstructured to completely structured data, at different levels of granularity and depth by allowing to trade-off depth for efficiency as required. The success of such a platform will largely depend on the integration of crosslingual technologies to increase the social reach and approach cross-cultural understanding.

3.2.3 E-Government

- Pan-European cross-border exchange of electronic documents, cross-border communication including legal aspects, specialised free translation services.
- Goals:
    - Creation of a multilingual Digital Single Market: we need cross-border public and government services that interoperate and counter market fragmentation, in particular in the areas of eGovernment, eProcurement and eHealth
    - Multilingual cross-border data value chains
    - E-procurement platforms where MT/LT can support translation of user interfaces, documents and large narratives that are currently performed manually; concept extraction; matching offer and demand to identify business opportunities and to produce accurate summaries for decision making in tendering.
- Means:
    - Development of terminologies, linked data sets, ontologies etc. that harmonise the concepts used in different countries and jurisdictions, as a basis to reach interoperability and develop a new generation of (public) services that is implemented across countries with multilingual technologies built in.
    - 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. Simplifying access to data by appropriate interfaces, e.g., based on natural language, is a crucial goal to achieve.
    - Generating reports and reviews automatically from data: It has been estimated that, in five years' time, more data will be generated automatically by machines than by humans. Although much of this content will be low-value advertising or journalism, an increasing proportion of it within the enterprise, hospital, government department and science laboratory will consist of highly actionable summary and review information. Language technology processes can take raw data and learn from precedents how to transform the numbers and words into succinct reports for later use by specialists will save time and money and rapidly inform all stakeholders for further discussion.

3.2.4 Online Dispute Resolution

- This solution relates to monolingual, cross-lingual, multilingual components for the Online Dispute Resolution initiative by ISA (Interoperability Solutions for European Public Administrations).

The solution foresees a multilingual platform to support 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 (especially at cross-border disputes), 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 MT 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. Thus, 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, but 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); thus 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.

3.3 Technology Solutions for Societal Challenges

The importance of languages for 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 works have shown, e.g., that even our moral decisions are influenced by whether we are speaking our mother tongue or a foreign language (Costa et. al 2014).

The solutions described in this section including those in the previous sections (see the intersection and overlaps of the different solutions) address many of the societal challenges specifically to be taken into account by activities under the framework of Horizon 2020.[20]

- Health, demographic change and wellbeing;
- Food security, sustainable agriculture and forestry, marine and maritime and inland water research, and the Bioeconomy;
- Secure, clean and efficient energy;
- Smart, green and integrated transport;
- Climate action, environment, resource efficiency and raw materials;
- Europe in a changing world - inclusive, innovative and reflective societies;
- Secure societies - protecting freedom and security of Europe and its citizens.

		Technology Solutions addressing Societal Challenges				
		Adaptable Interfaces for All	E-Health	E-Learning	E-Participation (from above)	BGCtoBGC (from above)
Societal Challenges	Health, demographic change, wellbeing	X	X	X		
	Food security, sustainable agriculture and forestry, marine and maritime and inland water research etc.					X
	Secure, clean and efficient energy				X	

[20] http://ec.europa.eu/programmes/horizon2020/en/h2020-section/societal-challenges

	Smart, green and integrated transport	X				
	Climate action, environment, resource efficiency and raw materials				(X)	X
	Europe in a changing world – inclusive, innovative and reflective societies	X		X	X	
	Secure societies – protecting freedom and security of Europe and its citizens	X				

(N.B. This table is a first attempt at mapping our solutions to the societal challenges; it's by no means meant to be final or complete.)

3.3.1 Adaptable Interfaces for All

- These solutions include assistive and accessible technologies for senior citizens or citizens with special needs, including speech and multi-modal interfaces, augmented reality, easy language, summarisation.

3.3.2 E-Health

- Cross-border healthcare scenarios with an emphasis on multilingual technologies.

3.3.3 E-Learning

- Combination of life-long learning with multilingual technologies, help immigration, cross-border migration, training for staff members of pan-european companies etc.

4 Enabling Platforms, Infrastructures and Services

We argue for and recommend the design and implementation of several ambitious large-scale platforms, infrastructures and services as turbo engines for research, innovation and development towards providing ubiquitous resources for the multilingual DSM and European society. The platforms will be used for fast and economical service delivery to enterprises and end-users as well as for testing, showcasing, proof-of-concept demonstration, avant-garde adoption, experimental and operational service composition.

Platforms help to reduce complexity on the side of users (in this case both end-users but also companies and other organisations that build new platforms on top of these services) and support evolution (competition and cross-fertilisation) on the side of the service providers. The concepts of hybrid research or DevOps, i.e., a tight loop of research, development and operations that allows for early testing and short development cycles, that has been successful in other areas will be adopted to language technologies, too. We envisage at least the following four platforms, infrastructures or sets of services that may or may not share basic services (maintenance, promotion, licensing, payment) and means for quality assurance:

- Translingual Trusted Cloud Platform for Human and Machine Translation
- Meaning and Knowledge Infrastructure
- Natural Language Interaction Services
- Text Analytics and Production Services

The services and resources provided by these infrastructures allow to implement and supply the solutions described in Chapter 3 and act as a bridge to the underlying research themes described

in Chapter 5. The range of services includes low-level technologies such as part-of-speech tagging and high-level (combined) ones such as machine translation including special terminology and human post-editing, automatic generation of spoken usage instructions, or email classification by sentiment and enrichment with background information.

The creation of powerful cloud computing platforms for a wide range of services dealing with human language, knowledge and emotion will not only benefit the individual and corporate users of these technologies but also the providers. Large-scale ICT infrastructures and innovation clusters such as this suggested platform are also foreseen in the Digital Agenda for Europe.

Users will be able to receive customised integrated services without having to install, combine, support and maintain the 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. Providers of language services rendered by human language professionals will be able to use the platform for enhancing their services by means of appropriate technology and for providing their services stand-alone or integrated into other application services.

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 in the platforms 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.

Citizens and corporate users will enjoy the benefits of language technology early and at no or reasonable costs through a large variety of generic and specialised services offered at a single source.

In order to allow for the gigantic range of foreseeable and currently not yet foreseeable solutions, the infrastructures and 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 can play an important role in the design of the platform.

Funding and business model of the platforms. The creation of these platforms has to be supported by public funding. Because of the high requirements concerning performance, reliability, user support, scalability, persistence as well as data protection and compliance with privacy regulation, the platforms need to be established by one or more consortia with strong commercial partners and also be operated by these consortia or commercial contractors. A similar platform with slightly different desiderata and functionalities is currently built under the name Helix-Nebula for the Earth Sciences with the help of the following commercial partners: Atos, Capgemini, CloudSigma, Interoute, Logica, Orange Business Services, SAP, SixSq, Telefonica, Terradue, ales, The Server Labs and T-Systems. Partners are also the Cloud Security Alliance, the OpenNebula Project and the European Grid Infrastructure. These are working together with major research centres in the Earth Sciences to establish the targeted federated and secure high-performance computing cloud platform.

The infrastructures and platforms are intended for a mix of commercial and noncommercial services. They would be cost-free for all providers of non-commercial services (cost-free and advertisement-free) including research systems, experimental services and freely shared resources but would raise revenues by charging a proportional commission on all commercially provided services. In order to reduce dependence on individual companies and software products, the base technology should be supplied by open toolkits and standards.

The infrastructures and platforms will considerably lower the barrier for market entry for innovative technologies, especially for products and services offered by SMEs. Still, these stakeholders may not have the resources, expertise, and time to create the necessary interfaces to integrate their

results into real-life services, let alone the overarching platforms themselves. There is still a gap between research prototypes and products that have been engineered and tested for robust applications. Moreover, many innovative developments require access to special kinds of language resources such as recordings of spoken commands to smartphones, which are difficult to get for several reasons.

The service platforms and infrastructures will be an important instrument for supporting the entire innovation chain, but, in addition, interoperability standards, interfacing tools, middle-ware, and reference service architectures need to be developed and constantly adapted. Many of these may not be generic enough to serve all application areas, so that much of the work in resource and service integration will have to take place.

4.1 Translingual Trusted Cloud Platform for Human and Machine Translation

One of our key goals is a multilingual European society and multilingual Digital Single Market, in which all citizens can use any service, access all knowledge, enjoy all media and control any technology in their mother tongues. This will be a world in which written and spoken communication is no longer hindered anymore by language barriers and in which even specialised HQ translation will be affordable.

Citizens, professionals, public organisations, companies or software applications in need of cross-lingual communication will use simple access points for channelling text or speech through a gateway that will instantly return the translations into the requested languages in the required quality and desired format.

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.

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. Based on the new technology, language-transparent web and language-transparent media will help realise a truly multilingual mode of online and media interaction for every citizen regardless of age, education, profession, cultural background, language proficiency or technical skills. 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

4.1.1 Implementation

If high-quality services are to be delivered right from the start, they will require a combination of Human Translation (HT), Computer-Assisted Translation (CAT) and full Machine Translation (MT). The envisioned platform will combine both free and fee-based services, including pre- and post-editing, language checking, etc. in addition to translation. Core requirements are trust in the reliability and accuracy of translation and the security of the translation channel. The platform should include:

- Quality-certified 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.)
- A combination of HT, CAT and fully automatic MT
- Extended services like multilingual text authoring, multimedia translation, and quality assurance by experts
- Mechanisms for customer care and trust building
- Certified security systems to ensure that confidentiality, privacy and data protection are designed into the platform from the outset
- Quality upscale models: services permitting instant quality upgrades if the results of the requested service levels do not yet fulfill the quality requirements
- Translingual spaces: dedicated locations for ambient interpretation. Meeting rooms equipped with acoustic technology for accurate directed sound sensoring and emission
- Mechanisms that allow the exploitation of nearly all usage data produced through human translation and post-editing (including use of scrambled/anonymised confidential data) for service/technology improvement
- A test bed and application target for MT research.

The last points are important as the platform is envisioned as a means to fulfill the high demand for collection of real usage data that is needed to improve MT performance. The platform is envisaged as a platform for LSPs as users and providers with two service layers:

Open Service Layer for the public

- 24/7 public service of automatic services (text and media translation, term and word translation)
- 24/7 public service for paid translation based on a bidding process

Provider Service Layer for providers

- Free and paid 24/7 services MT, terminology, dictionaries, language checking, TMs, quality assurance, etc. for participating LSPs (and technology providers)

The free services could include:

- MT services provided by private companies
- MT services provided by EC (maybe restricted)
- MT services by systems from research centers
- HT/postediting services for special purposes provided by NGO-initiatives, e.g., Translators without Borders, Rosetta Foundation
- HT/postediting services in special situations (by crowdsourcing to registered screened volunteers)

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 source data and translations.

There needs to be a close cooperation scheme between language industry (LSPs), translation technology industry (MT terminology, translation process and management systems, language checking) and MT research. A scheme for quality assurance including accepted shared quality metric for automatic, human and hybrid translation (MQM, TAUS DQF) and tools and processes for human and partially automated quality assessment are mandatory.

The technical solutions will benefit from new trends in IT such as software as a service, cloud computing, linked open data and semantic web, social networks, crowdsourcing etc. For MT, a combination of translation brokering on a large scale and translation on demand is promising. The idea is to streamline the translation process so that it becomes simpler to use and more

transparent for the end user, and at the same time respects important factors such as subject domain, language, style, genre, corporate requirements, usability and user preferences. Use of translations must be fully instrumented so that its efficacy can be continuously assessed and solution adapted to changing language usage requirements. Technically, what is required is maximum interoperability of all components (corpora, processing tools, terminology, knowledge, maybe even translation models) and a cloud or server/service farm of specialised language technology services for different needs (text and media types, domains, etc.) offered by SMEs, large companies or research centres.

Solutions for better communication and for access to content in the users' native languages would reaffirm the role of the EC to serve the needs of the EU citizens. A connection to the infrastructure programme 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 in which the European social and societal needs massively overlap with business opportunities to achieve funding investment that pays back, ideally public-private partnerships. Concerted activities sharing resources such as error corpora or test suites and challenges or shared tasks in carefully selected areas should be offered to accelerate innovation breakthrough and market-readiness for urgently needed technologies.

4.2 Multilingual Meaning and Knowledge Infrastructure

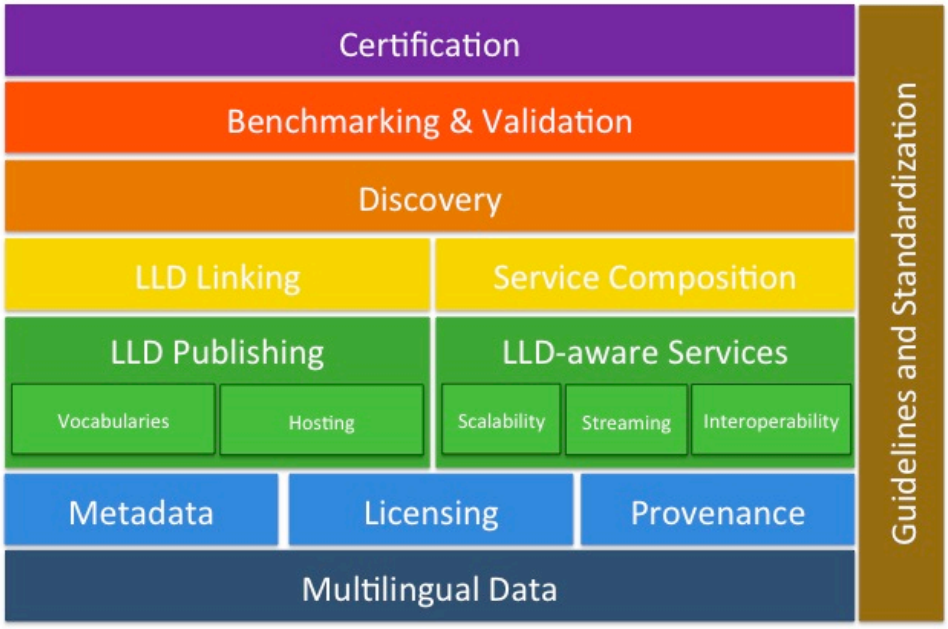
(Includes LOD, knowledge repositories and graphs, data economy, generic data value chains etc.; certain advanced/avantgarde functionalities still need to be included here)

This infrastructure will bring in services for processing and storing knowledge gained by and used for understanding and communication. It will include repositories of linked data and ontologies, as well as services for building, using and maintaining them. These in turn permit a certain range of rational capabilities often attributed to a notion of intelligence. The goal is not to model the entire human intelligence but rather to realise selected forms of inference that are needed for utilising and extending knowledge, for understanding and for successful communication. These forms of inference permit better decision support, pro-active planning and autonomous adaptation. A final part of services will be dedicated to human emotion. Since people are largely guided by their emotions and strongly affected by the emotions of others, truly user-centred IT need facilities for detecting and interpreting emotion and even for expressing emotional states in communication.

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, corpora and online content and data sets of all types. Extensive sets of proven open source tools exist and there is a rapid migration of language resources to this technological platform.

The three core principles of the platform are:

1. Linked Data ensures that data and services form a linked ecosystem rather than a set of fragmented and non-interoperable datasets and services. A growing set of standardised linked data vocabularies ensure convergence.
2. Semantic Technologies conformant to web standards such as RDF and OWL offer powerful APIs such as SPARQL for search and RESTful services to publish, update and manipulate linked data on the web.
3. De-centralization 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 platform referred to collectively as **Linguistic Linked Data** (LLD) and is structured as follows: **Multilingual data**, in all forms, modality and media types form substrate of the platform. Mappings to existing common data format such as XML vocabularies, JSON and CSV ensure the major benefits of the platform are gained without costly transformation of existing data. **Metadata:** providing basic information about the dataset (author, language, structure), etc. **Licensing:** specifying the terms and conditions of use of linguistic resources should be specified. This includes the description of copyright information and any other rights-related restriction (e.g., privacy and data protection of personal data and commercial paywall access control where needed). **Provenance:** describing the origin and processing history of data, which is key to assessing its usability in a specific task

The LLD-specific layers provide the integration point with the other foreseen platforms, infrastructures and services and comprises of the following two layers:

- **Linguistic Linked Data Publishing** consists of guidelines, best practices and standards describing how different types of resources (lexica, corpora, terminologies, lexico-semantic resources) should be transformed into RDF and how they should be published on the Linguistic Linked Open Data (LLOD) cloud. This layer also comprises of concrete tools and frameworks supporting translformation and publication as well as recommendation on use of common vocabularies and data hosting.
- **Linguistic Linked Data Linking**: This layer comprises of guidelines, best practices, tools and frameworks to supporting linking of resources as well as concrete tools and frameworks to support semi-automatic linking of resources, including managing links between resources with different access terms and conditions.

For **LLD Services**, the following two layers are included:

- LLOD-aware Services: Addresses content and knowledge analytics and processing services that can consume and produce LLD. This includes:
    - **Scalability:** LLOD-aware services should be able to scale to processing large amounts of data. This requires a non-centralized architecture in which services can cache results and pass intermediate results to other services instead of relying on a client to coordinate and interact with all services implementing a complex workflow.
    - **Streaming:** The architecture relies on streaming principles to support the implementation of services that can process data in a stream fashion, thus reducing overhead of creating and closing connections, supporting real time analytics.
    - **Interoperability** focussed on use of common vocabularies to describe data inputs

and output of services.

- **Service Composition** for the chaining of single LLD-aware services to implement, monitor and optimise more complex workflows combining NLP, data management and human elements.

Discovery of LLD data set implemented by an arbitrary number of services that index and aggregate datasets and services and expose a standardized API that supports querying a repository to find datasets that meet certain criteria. The tools in the discovery layer should support SPARQL but provide Linked Data that is both understandable and searchable by both humans and machines.

Benchmarking and Validation to support the comparison of datasets and services to allow potential users to choose the service or dataset that best meets their needs using a common set of tools and quality definitions.

Certification to allow independent services and agents to assign quality labels or certificates to datasets if they meet specified conditions.

Guidelines and Standardisation is orthogonal to the above mentioned layers and emphasizes that standardisation and promotion of uptake by appropriate community initiatives is crucial to ensure wide acceptance, implementation and use of LLD Platform.

4.2.1 Implementation

Many elements of the platform are already in place, based on general 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 communities including the Linked Data for Language Technology Community, the best Practices in Multilingual Linked Open Data community, the Ontological-Lexical Community, the Open Linguistic Group at the Open Knowledge Foundation and the Sentiment Data community. Massively multilingual examples of aggregation and discovery solutions using LLD are publically available and already having a major impact on the NLP and language resource communities. Babelnet[21] aggregates public lexical-conceptual information from Wikipedia, Wikidata, different Wordnets and Wiktionary into a single discovery service supported by the text 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. Linghub[22] is a metadata aggregator for language resources. Starting from the migration of META-SHARE meta-data to RDF, this service now integrates metadata from 100,000 resources from across 76 existing repositories including META-SHARE, CLARIN, LREMap, and LEXVO, to name a few. The Linguistic Linked Data Cloud[23] of interlinked language resources is now an important and growing part of the overall Linked Open Data cloud. The language resource community is therefore well on the way to wholesale adoption of linked data as its primary data exchange mechanism.

4.3 Natural Language Interaction Services

(Includes all NLI/HLT/spoken services, both ASR, TTS, dialogue systems, conversational technologies, multilingual via bridge to Translingual Trusted Cloud etc.)

Both basic language processing and understanding will be used by services that support human communication or realise human-machine interaction. Part of this layer are question answering and dialogue systems as well as email response applications.

4.4 Text Analytics and Production Services

(Includes summarisation, opinion mining, trend mining, sentiment analysis, POS tagging,

[21] http://babelnet.org/
[22] http://linghub.org/
[23] http://linguistic-lod.org/llod-cloud

dependency parsing, information extraction, document enrichment through bridge to Meaning and Knowledge, multilingual through bridge to Translingual Trusted Cloud, content curation, content production etc.)

A top layer consists of language processing such as text filters, tokenisation, spell, grammar and style checking, hyphenation, lemmatising and parsing. At a slightly deeper level, services will be offered that realise some degree and form of language understanding including entity and event extraction, and opinion mining.

5 Research Themes

Although we use computers to write, telephones to chat and search the web for knowledge, IT has no direct access to the meaning, purpose and sentiment behind our trillions of written and spoken words. This is why today's technology is unable to summarise a text, answer a question, respond to a letter or to translate reliably, let alone to implement some of the more complex solutions envisaged above.

Many companies had started much too early to invest in language technology research and development and then lost faith after a long period without any tangible progress. During the years of apparent technological standstill, however, research continued to conquer new ground. The results are a deeper theoretical understanding of language, better machine-readable dictionaries, thesauri and grammars, specialised efficient language processing algorithms, hardware with increased computing power and storage capacities for big data, large volumes of digitised text and speech data and new methods of statistical language processing that could exploit language data for learning hidden semantic regularities governing our language use.

We do not yet possess the complete know-how for unleashing the full potential of language technology as essential research results are still missing, but the speed of research keeps increasing and even small improvements can already be exploited for innovative products and services that are commercially viable. We are witnessing a chain of new products for a variety of applications entering the market in rapid succession. These applications tend to be built on dedicated computational models of language processing that are specialised for a certain task.

But increasingly we observe a reuse of core components and language models for a wide variety of purposes. It started with dictionaries, spell checkers and text-to-speech tools. Google Translate, Apple's Siri, Microsoft's Cortana and IBM Watson still do not use the same technologies for analysing and producing language, because the generic processing components are simply not powerful enough to meet their respective needs. But many advanced research systems already utilise the same tools for syntactic analysis. This consolidation process is going to continue. In ten years or less, basic language proficiency is going to be an integral component of any advanced IT.

In the envisaged big push toward realising the solutions sketched in the previous chapters by massive research and innovation, our community is faced with three enormous challenges:

1. Richness and diversity. A serious challenge is the sheer number of languages, some closely related, others distantly apart. Within a language, technology has to deal with numerous dialects, sociolects, registers, professional jargons, genres and slangs.
2. Depth and meaning. Understanding language is a complex process. Human language is not only the key to knowledge and thought, it also cannot be interpreted without certain shared knowledge and active inference. Computational language proficiency needs semantic technologies.
3. Multimodality and grounding. Human language is embedded in our daily activities. It is combined with other modes and media of communication. It is affected by beliefs, desires, intentions and emotions and it affects all of these. Successful interactive language technology requires models of embodied and adaptive human interaction with people, technology and other parts of the world.

It is fortunate for research and economy that the only way to effectively tackle the three challenges involves submitting the evolving technology continuously to the growing demands and practical stress tests of real world applications. Google's Translate, Apple's Siri, Microsoft's Cortana, Autonomy's text analytics and scores of other products demonstrate that there are plenty of commercially viable applications for imperfect technologies. Only a continuous stream of technological innovation can provide the economic pull forces and the evolutionary environments for the realisation of the grand vision.

In the following sections, we will briefly present the main envisaged research approaches and solutions together with targeted breakthroughs within four major research themes:

- Research Theme 1: HQ Machine Translation and Human Translation
- Research Theme 2: Crosslingual and Multilingual Big Data Text and Speech Analytics
- Research Theme 3: Conversational Technologies and Natural Language Interfaces
- Research Theme 4: Meaning and Knowledge

5.1 Research Theme 1: HQ Machine Translation and Human 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. Yet, many more resources and a more fundamental, novel scientific approach – that eventually runs across several projects and also calls – are needed for significant and substantial improvements that cover the phenomena and problems that make up the Zipf'ian long tail. This is an obstacle in particular for individual research centres and SMEs given their limited resources and planning horizon.

5.1.1 Novel Research Approaches and Targeted Breakthroughs

Although recent progress in MT has already led to many new applications of this technology, radically different approaches are needed to accomplish the ambitious goal of this research including a true quality breakthrough. Among these new research approaches are:

- More focus on high-quality, publishable outbound-translation that is needed for the success of MT in the language industry[24]
- Systematic concentration on quality barriers, i.e., on obstacles for high quality
- A unified, dynamic-depth weighted, and multidimensional quality assessment model with task and language profiling
- Strongly improved automatic quality estimation for given task specifications
- Inclusion of translation professionals and enterprises in the entire research and innovation process (plus Inclusion of technologists into research on human translation processes)
- Ergonomic work environments for computer-supported creative top-quality human translation and multilingual text authoring
- Improved statistical models that extract more dependencies from the data
- Discriminative and integrated training, integrating search and translation process into training
- A Semantic translation paradigm by extending statistical translation with semantic data such as linked open data, ontologies including semantic models of processes and textual inference models
- Stronger emphasis on the properties of individual languages: Exploitation of strong monolingual analysis and generation methods and resources
- Modular combinations of specialised analysis, generation and transfer models, permitting

[24] 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.

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
- Seemingly creative translation skills by analogy-driven transfer models
- Automatic subtitling and voiceover of films and multimedia applications in selected domains, such as public service (parliament recordings, legal proceedings), sports events, and other applications (TV archives, movies, online services at content providers);
- Ambient translation
- Always-correct translation for critical subdomains

5.1.2 Solution and Realisation

Cooperation with translation professionals. A close cooperation of language technology and professional language services is planned. Professional translators and post-editors are required whose judgements and corrections will provide insights for a more analytical and systematic approach of quality boundaries and data for bootstrapping new methods. The cooperation scheme of research, commercial services and commercial translation technology is planned as a symbiosis since language service professionals or advanced students in translation studies or related programmes working with and for the developing technology will at the same time be the first test users analytically monitored by the evaluation schemes. This symbiosis will lead to a better interplay of research and innovation.

Novel quality metrics and human annotation. The translation model improvements needed for HQMT need to be based on novel reliable and informative quality measures since common measures such as BLEU or edit-distance based measures such as TER may incorrectly punish perfectly good translations which differ from a given reference (or references), e.g., in completely legitimate word order and/or morphological realisation. Currently, the only way of assessing translation quality on the needed level of reliability and granularity (word/phrase level) involves manual work such as post-editing or explicit error annotation to be collected in cooperation with LSPs, Shared Tasks, etc. This data is needed for system development (generating research hypothesis, driving development tasks) and as test case for testing the performance of new models using advanced diagnostic tools. The mid-term goal is to automate novel metrics as far as possible including sampling functionality; incorporate feedback from research systems, develop datasets for new metrics and best practices.

In the new type of MT development, annotation is performed on one of the three levels as needed:

1. Phenomenological level. This comprised markup of issues on the translated output (target side) using an annotation metric like TAUS DQF or MQM developed in the QTLaunchPad project.

2. Linguistic level. This comprises markup on the translation source or target side with information like part of speech, phrase boundaries or more specific phenomena under consideration such as long-distance dependencies or multi word expressions.

3. Explanatory level. This comprises markup of the source (with hindsight to the target) with (typically speculative) reasons for translation failure such as model class, n-gram size, data sparseness, etc.

The annotation on the phenomenological level usually involves language professionals like human translators while the other two levels require linguistic skills and expertise on MT system level that researchers from linguistics, LT, and related areas typically have.

Translation quality is always relative to the given purpose and circumstances that are ideally captured in a formal specification. Together with industry, translation error metrics and assessment frameworks should be standardised to support comparison and compatibility of results.

Apart from direct error annotation, a recent idea is to exploit translation output that was corrected by human translators through post-editing to improve learning-based MT. This output should be used to update, reinforce, and correct systems' translation hypotheses and together with explicit error markup will help to overcome real barriers and also fix relatively minor issues like punctuation or agreement that seem to have been overlooked in the development of MT engines for gisting, yet render most output improper for outbound translation.

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

Another use of the novel quality estimation methods and metrics that can accurately predict errors in previously unseen texts at various levels of granularity ist to define source and target-language "profiles" for different problems that may arise for various language pairs and system types. The profiles will help MT developers dealing with sparse data in under-resourced languages by giving them guidance on how to locate particularly relevant examples in corpora or by highlighting particularly important issues to consider in developing hybrid systems.

Evaluation and development cycle. A typical evaluation and development cycle might look as follows. Given some MT translated corpus and initial hypotheses of what issues may be encountered, the following steps are included:

1. Definition of a concrete quality metric for the given purpose starting from an existing metric of from scratch using a metric builder.
2. Filtering/sampling the translation corpus to be evaluated in a triage: (a) Perfect translations. (b) Almost good translations that need further analysis, (c) Bad translations that do not qualify for further use. These steps can be performed manually, supported by a very basic score card or performed in a semi-automatic or even automatic way using a quality estimation toolkit depending on the size of the corpus, available human resources, and required precision and recall.
3. Deeper analysis of the segments of type b. If information on the segment level is sufficient, a score card can be used, for detailed error annotation, annotation tools like translate5 can be used.
4. (Re-)Training of quality estimation tools on the newly annotated data from step 3 and possibly on the new filtered data from step 2.
5. Inspection of the errors to: (a) Confirm if the system output supports the hypotheses, (b) Get a quantitative basis to decide on MT development priorities, (c) get a qualitative idea of remaining quality barriers

Based on the insights gained in step 5 (and the perfect translations gained in step 2 if the number is large enough), the MT engine can then be improved. If the metric needs to be adjusted, another

development cycle starts at step 1, otherwise, new translations of an improved engine can feed into step 2.

Through the improvement of quality estimation over time, automation of the manual steps should become more reliable and reduce the human resources needed.

Better and more semantic models. Many language pairs are difficult for current string-based translation and reordering models. Complex structural and lexical mappings are often not inferrable from the linguistic surface even from huge amounts of training data. As a result we need better translation models, more semantic abstraction, and generalisation. Improved models will be better able to bridge between diverse languages (word order, long range phenomena) and deliver higher-quality translations. Challenges are that complex representations need complex and potentially computationally costly decoding and require annotated training material for supervised methods. A huge potential lies in the usage of Linked Open Data where lots of parallel or comparable semantically annotated data exists. Machine translation should make use of both the informatin contained in the knowledge representations and in the linked documents to fill the "knowledge gap". This new approach will require graph rather than tree representations, and might make use of recent developments in deep learning for learning of synchronous dependency/semantic representations on the right level of granularity to connect source/target syntactic/semantic representations. These then have to feed into novel, efficient decoders for complex syntactic and semantic models.

Platform for MT research and development. The procedure outlined above pertains to both MT development in research and production context. It should be tested and further developed into more standardized pipelines. A large-scale evaluation infrastructure, structured to areas, applications, and languages has to be designed and implemented for the resource and evaluation demands of large-scale collaborative MT research. An initial inventory of language 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 third-party (customer) data is needed for realistic development and evaluation, intellectual property rights and legal issues must be taken into account from the onset. The infrastructures to be built include service clouds with trusted service centres, interfaces for services (APIs), workbenches for creative translations, novel translation workflows (and improved links to content production and authoring) and showcases such as ambient and embedded translation.

Innovation Projects should "stress test" these technologies in realistic scenarios. This level of validation is vital since success in shared tasks (e.g., WMT) may not equate to usefulness in a real production environment, where concerns of stability, maintenance, efficiency, or adaptability may be a deciding factor

Although the research strand will focus on advances in translation technology for innovation in the language and translation service sector, a number of other science, technology and service areas need to be integrated into the research from day one. Some technology areas such as speech technologies, language checking, authoring systems, analytics, generation and content management systems need to be represented by providers of state-of-theart commercial products.

5.2 Research Theme 2: Crosslingual and Multilingual Big Data Text and Speech Analytics

The research topics in this theme will provide breakthroughs in the area of Big Data Text and Speech Analytics. Applications of these research results will, among others, provide businesses to adapt and communicate with their customers. It will increase transparency in decision-making processes, not only in business and society but also in politics. Powerful analytical methods 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 so that they can react quickly to new trends and optimise their marketing and customer communication strategies.
- Companies will get the desperately needed instruments to exploit the knowledge and expertise of their huge and diverse workforces, the wisdom of their own crowds, which are the most highly motivated and most closely affected crowds.
- Companies will be able to adapt to new geographical and cultural contexts and realize international investments with low risk.
- Political decision makers will be able to analyse public deliberation and opinion formation processes in order to react swiftly to ongoing debates or important, sometimes unforeseen events.
- Citizens will be able to access multicultural, multilingual and from multiple political perspectives information, which will lead to the consumption of validated and non-contradictory information and will reduce instability and insecurity in Europe.
- Citizens and customers get the opportunity (and necessary information) to participate and influence political, economic and strategic decisions of governments and companies, ultimately leading to more transparency of decision processes.

Thus, leveraging collective and social intelligence in developing new solutions to these 21st century challenges seems a promising approach in such domains where the complexity of the issues under discussion is beyond the purview of single individuals or groups.

The research and innovation will provide technological support for emerging 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. At the same time the resulting technologies will be applicable to smaller groups and also interpersonal communication as well, even though different dynamics of information exchange can be foreseen. The research to be carried out and technologies to be developed in this priority theme will also 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. What we learn from processing language is the prime tool for processing the huge and intractable data streams that we will be confronted with in the near future.

5.2.1 Novel Research Approaches and Targeted Breakthroughs

A key enabler will be language technologies that can map large, heterogeneous, and, to a large extent, unstructured volumes of online content to actionable representations that support decision making and analytics tasks. Such mappings can range from the relatively shallow to the relatively deep, encompassing for example 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, data mining, automatic linking, content validation, reasoning and summarisation have to be made interoperable with knowledge representation and semantic web methods such as ontological engineering (cf. Meaning and Knowledge theme). Drawing expertise from related areas such as knowledge management, information sciences, or social sciences is a prerequisite to meet the challenge of modelling social intelligence. The new research approach should target the bottleneck of knowledge engineering by:

- Semantification of the web: bridging between the semantic parts and islands of the web and the traditional web containing unstructured data;
- Multidimensional integration of textual and multimedia data with social network and social media data; these dimensions include semantics, context, location and especially the temporal;
    - Common representations that can deal with the requirements imposed by data produced nowadays, i.e., big, heterogeneous, distributed, user-tagged, user-generated,

multimodal; such a representation has to be light-weight and empowered with semantics.
    - Representation alignment methods
- Aligning and making comparable different genres of content like mainstream-news, social media (blogs, twitter, facebook etc.), academic texts, archives etc.;
- Extracting semantic representations from social media content, i.e., creating representations for reasoning and inferencing;
- Taking metadata and multimedia data into account. The following list contains specific targeted breakthroughs to be sought in this scenario:
    - Social intelligence by detecting and monitoring opinions, demands, needs and problems;
    - Detecting diversity of views, biases along different dimensions (e.g., demographic) etc. including temporal dimension (i.e., modelling evolution of opinions);
    - Support for both decision makers and participants;
    - Problem mining and problem solving ;
    - Support of collective deliberation and collective knowledge accumulation;
    - Vastly improved approaches to sentiment detection and sentiment scoring (going beyond the approach that relies on a list of positive and negative keywords);
    - Introducing genre-driven text and language processing (different genres need to be processed differently);
    - Personalised recommendations of e-participation topics to citizens;
    - Proactive involvement ne-participation activities;
    - Understanding influence diffusion across social media (identifying drivers of opinion spreading);
- Retrieval techniques from heterogeneous resources.
    - Passage retrieval to support question-answering tasks from heterogeneous content including social media.
    - Query rewriting methods
    - Resource selection to support search in federated and distributed environments
- More sophisticated methods for topic and event detection that are tightly integrated with the Semantic Web and Linked Open Data.
    - Multimodal clustering approaches based on heterogeneous features; Automatic and user-driven clustering; clustering and classification based on semantic representations.
    - Event detection in multimedia by exploiting semantic and textual features from speech recognition and captions, as well as visual and motion information.
- Modelling content and opinion flows across social networks;
- Speech recognition in noisy environments
- Decision support techniques
    - Based on semantic reasoning (rule based, fuzzy, backward and forward chaining) techniques that operate on semantically integrated data
    - Based on supervised models trained with a variance of features (e.g. concepts, name entities, n-grams, contextual chacteristics, sentiment)
    - Based on visual analytics that provide an interpretable version of the aforementioned techniques.
- Cross validation of content in media and social media;
    - Detection of fake content
    - Identification of contradictory facts
    - Identification of hidden relations including repeated/similar facts along the spatiotemporal axes.
- Summarisation of content
    - Multidocument summarization
    - Extractive (shallow) summarization
    - Abstractive (semantic relation-driven) summarization of multimodal documents

- Evaluation of methods by analytic/quantitative and sociological/qualitative means.

5.2.2 Solution and Realisation

Individual solutions should be assembled from a repository of generic monolingual and cross-lingual language technologies, packaging state-of-the-art techniques in robust, scalable, interoperable, and adaptable components that are deployed across sub-tasks and sub-projects, as well as across languages where applicable (e.g., when the implementation of a standard data-driven technique can be trained for individual languages). These methods need to be combined with powerful analytical approaches that can aggregate all relevant data to support analytic 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 requirements of their applications. Technology should be able to work with heterogeneous sources, ranging from unstructured (arbitrary text documents of any genre) to structured (ontologies, linked open data, databases).

To accomplish interoperability we suggest a strong 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 (be it lexical, propositional, or referential) and extralinguistic semantics (e.g., taxonomic world or domain knowledge). For example, grammatical analysis (which one may or may not decompose further into tagging, syntactic parsing, and semantic role labelling) should make available a sufficiently abstract, normalised, and detailed output, so that downstream processing can be accomplished without further recourse to knowledge about syntax. Likewise, event extraction or fine-grained, utterance-level opinion mining should operate in terms of formally interpretable representations that support notions of entailment and, ultimately, inference.

Finally, 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 (e.g., part-of-speech taggers) may need to be customised or re-trained to deliver satisfactory performance. Further taking into account variation across downstream tasks, web-scale 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 tradeoffs need to be documented empirically, and component technologies accompanied with methods and tools for adaptation and cost-efficient re-training, preferably in semi- and un-supervised settings.

The technical solutions needed include:

1. Technologies for decision support, collective deliberation and e-participation.
2. A large public discussion platform for Europe-wide deliberation on pressing issues such as energy policies, financial system, migration, natural disasters, etc.
3. Visualisation of social intelligence-related data and processes for decision support (for politicians, health providers, journalists, manufacturers, entrepreneurs or citizens).
4. High-throughput, web-scale content analysis techniques that can process multiple different 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.
5. Mining e-participation content for recommendations, summarisation and proactive engagement of less active parts of population.
6. Detection and prediction of events and trends from content and social media networks.

7. Extraction of knowledge and semantic integration of social content with sensory data and mobile devices (in near-real-time).
8. Cross-lingual technology to increase the social reach and approach cross-culture understanding.

5.3 Research Theme 3: Conversational Technologies and Natural Language Interfaces

5.4 Research Theme 4: Meaning and Knowledge

(This includes Semantics, Knowledge Representation, Inferencing, User Modelling, Semantic Web, Linked Data: Rich semantic user profiles, semantics for objects, individuals, groups, intentions, contexts – including multi-modal, gestures, eye tracking, sensor data –, cultures, text types/genres, location, etc.)

The Digital Single Market aims to lower the barrier to cross-border commerce through wider broadband connectivity and common rules for data protection, licensing and online purchases. However, a common infrastructure is also needed to enable effective customer engagement across the many language barriers present within the EU and towards markets globally. Language Technologies enable companies to scale-up their ability to engage effectively with customers in different target market languages. While Machine Translation is a key enabling technology, other language technologies are also needed to tailor engagement between companies and their customers to the domain being addressed. Customers should be able to search for user-generated content on a product or service regardless of the language in which it was originally posted. Image, video and audio postings on products should also be tagged, summarised, discoverable and accessible to users in any other language, who in turn should be able to post back comments regardless of language barriers. 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 customer experiences, companies must closely monitor and analyse user-generated content on third-party social media, blogs, question-answer forums and product review sites and react continuously with well targeted customer engagement. It is only when Language Technology enables the communities of online customers interested in a product to interact seamlessly across languages that the full potential of adaptive, responsive customer engagement will be realised.

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. Therefore, one-size-fits-all language technologies, such as free machine translation, will fail to meet the language needs of specialised SMEs.

SMEs however typically lack the knowledge or technical capacity to assemble their own linguistic data assets and use them to tailor language technology to their needs. However, without tailored language technology support, SMEs will not be able to avail of the DSM because of the language barriers to bidirectional customer engagement, even if other communication and regulatory barriers are removed. Linguistic Linked Data is already proving a scaleable source of massively multilingual open language resources for LT services. However research is needed into tools and techniques to integrate the lifecycle management of linguistic data into language 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 in the wider customer engagement ecosystem.

To put this in context, multilingual content is growing at an impressive, exponential rate. Exabytes of new data are created every single day. 90% of the data available today has been generated in

the last two years only. The International Data Corporation (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.

Data has been recently referred to as the “new oil” of the digital economy. However, crude oil is useless unless it is refined. And the same holds for multilingual data: (i) If data is not linked to other data it can only be used in isolation, rather than in context; (ii) If data is not analysed further, then no insights can be generated; (iii) If data is not verified nor the provenance of data can be tracked, it cannot be trusted; and (iv) 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 and ultimately value to companies, public organizations and citizens alike. Only if such an ecosystem is available will the “new oil” generate added value. Otherwise, it will remain unexploited, unrefined and thus useless.

5.4.1 Novel Research Approaches and Targeted Breakthroughs

The main dimensions that need to be prioritised for investment of research and development are:

1. **Linking:** Only if data is linked across sources can datasets be exploited in context, making more of the single dataset compared to using it in isolation only. Data linking is thus crucial to exploit data, and investments in new methodologies for data linking are needed. As the amount of data grows, it will become harder and harder to find the data that is most appropriate to solve a particular task. We need to create an ecosystem that fosters also the discovery of relevant data, and linking plays a crucial role in this endeavour. After all, the current web search technology would not exist without links between pages! We need to create a similar ecosystem for multilingual data (Web of Data), where links become first class, value-add objects and tools are available to manage the relevance, authoritativeness and quality of links.
2. **Generating insights from unstructured data:** Most available data is in textual form or comes from sensors and is thus unstructured, 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, text mining and natural language processing methods play a crucial role here, but need to be extended in terms of coverage, robustness and scalability by significant investments.
3. **Trust and Usability:** For data to be exploited in applications, trust in the data is key. Trust involves on the one hand 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 thus remain attached to data over the whole data lifecycle from the creation of the data, use of the data through to its derivation and modification. This metadata must be available in machine-readable form so that applications can automatically process and ensure compliance with the licensing terms and ensure trust.
4. **Privacy and Data Protection:** An ecosystem of linked 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 a linked data 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. The ubiquity of electronic eyes and ears in the IoT that can then feed machine learning for LT presents massive new challenges in how users understand and control how their word, utterances and exceptions are used.

5. **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.
6. **Access to information and services without borders:** In the 21st century, we cannot afford that access to data and services stops at the borders of countries due to language barriers. Much in the same way that in Europe we have invested in the free flow of goods and people, we need to substantially invest into the cross-border flow of data and availability of public and commercial services but also in homogenization and consistency of services across borders and languages. This requires substantial investment into integrating semantic and linked data technologies with localization and human language technologies, in particular through the use of standards.

If Europe does not substantially invest in the above 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 on the Web and with what level of privacy. This clearly is a threat to the free flow of information and constitutes a modern form of information hegemony by big players that runs counter to the free and independent availability of information that is required to strengthen democracy that is key to the European tradition. The key values of decentralisation and subsidiarity of authority combined with multilingualism means Europe is ideally placed the lead the world in linguistic linked data and therefore in the accurate tailoring of adaptive LT that this enables.

5.4.2 Solution and Realisation

The encoding of knowledge seemed to be a promising alternative to the current web, so that the vision of the Semantic Web was born. Its main bottleneck, however, remains the problem of knowledge acquisition. The intellectual creation 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 also 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 management and communication in everyday tasks using natural language interfaces.

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. We will see wide use of automatic subtitling and first successful examples of automatic voice over for a few languages, but full and reliable interlinking of audio and video resources to parallel data is needed to achieve accurate solutions built on machine learning.

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 an understanding of the context in which the question is asked. Sensing and 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 search proactively and can make inferences from their (possibly 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 2024, we envisage such systems which operate on huge, dynamic, heterogeneous data streams, and which also provide powerful approaches to navigation and visualisation. In many cases it will be important to consider issues such as data provenance, trust, privacy, data protection, security, and rights. In particular, 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, in particular, relates to positive (democracy) and negative (surveillance) aspects of large-scale multimodal knowledge integration and access.

5.5 Core Resources and Technologies for Language Production and Analysis

An essential, important prerequisite for all infrastructure activities is pooling and sharing language resources and technologies. In this regard, one of our key goals is to set up a shared programme together with, ideally, all EU Member States and all interested associated countries, in order to collaborate closely with all research centres and universities in the different countries, 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 modeling and computational linguistics methods for automatic language processing and generation.

The four research themes 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, information retrieval tools, machine learning toolkits, speech recognition and speech synthesis engines, and integrated architectures such as GATE and UIMA.

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. Both tools and resources can be rather general or highly task- or domain-specific, tools can be language-independent, datasets are, by definition, language-specific. As complements to the core technologies and resources there are several types of resources, such as error-annotated corpora for machine translation or spoken dialogue corpora, that are specific to one or more of the four research themes.

A key component of this research and innovation agenda is to collect, develop and make available core technologies and resources through one or more shared infrastructures 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 technologies 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 no or almost no resources or basic technologies available.

The European academic and industrial technology community is fully aware of the need for sharing resources such as language data (e.g., corpora), language descriptions (e.g., lexicons, thesauri, grammars), tools (e.g., taggers, stemmers, tokenisers) and core technology components (e.g., morphological, syntactic, semantic processing) as a basis for the successful development and implementation of the priority themes. 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. All language resources and basic technologies will be created under the core technologies umbrella. The effort will revolve around the following axes: Infrastructure; Coverage, Quality, Adequacy; Language Resources Acquisition; Openness; Interoperability.

6 Horizontal Framework Aspects

6.1 Copyright and IPR

Research and innovation in Language Technology (LT) depends on language data the way climate research depends on weather data or economic studies depend on financial data. Results derived in LT research from the analysis of large amounts of texts in areas like Machine Translation, Text and Data Mining (TDM) 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 piece of artwork, document, etc. Still, traditional copyright and half-hearted exceptions for research are experienced as huge obstacles for research and innovation by the European research community. 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.

The current research exception in Art. 5.3 a) of the Directive 2001/29/EC allows Member States to adopt copyright exceptions for "non-commercial scientific research". For example, the UK has recently implemented a range of copyright exceptions, including for education purposes, and for research, private study and text and data analysis for non-commercial research. Ireland is at present considering similar exceptions. Recent reforms in Germany and Spain did not address this issue. Nor do proposed reforms in Finland. Moreover, the transposition of Art 5.3(a) is **not mandatory**, and therefore there are important differences between Member States as to the content of these exceptions.

Furthermore, in many Member States they are subject to additional safeguards (e.g., allowing only "small parts" of works to be used) or to a flat-rate payment to be made to a collecting society. Combined with unclear rules governing conflict of laws (which law to apply in a trans-border situation?) this is a serious obstacle to **international research projects**. It would be desirable that all languages in Europe can be researched under the same condition across all member countries.

Moreover, the notion of non-commercial scientific research lacks clarity. It is **not clear** if **applied research** (as opposed to fundamental/basic research), or research carried out in public-private partnerships can also be subject to the exception.

Finally, in most EU jurisdictions (but not in Ireland or the UK) copyright exceptions can be overridden by contracts (if a contract prohibits certain uses of works, they cannot be made, even if allowed by an exception). In a world where works are mostly used on a contractual basis (via copyright licenses), this can reduce a copyright exception to a meaningless clause. The same applies to technological protection measures (art. 6 of the Directive) – by applying such a measure, the right holder can de facto prevent users from making certain uses of the work (e.g., lawful private copies, or lawful copies made for research). The Court of Justice of the European Union (CJEU) is trying to solve the problem (in decisions such as Nintendo C-355/12 or VG Wort C-457/11 — C-460/11), but a robust **legal basis is still lacking**.

However, "modifying copyright rules to reflect new technologies" is a priority of the "Digital Single Market" plank of the EU Commission's 2015 Work Programme, so that robust legal basis may be forthcoming from that process. In that context, and without restricting the process of research to some particular field, we need a copyright exception that

- deems legal the transitional copies that happen in the course of TDM

- enables researchers to legally build corpora
- allows researchers to legally hand those to other researchers
- allows researchers to legally publish research results (as far as research results are based on mining of unpublished copyrighted material, the approval of the creators of such material should still be necessary).

We generally support the Draft Reporton the implementation of Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society issued by the Committee on Legal Affairs of the EP (2014/2256(INI))[25], in particular the passages cited below:

Exclusive rights

4. Considers the introduction of a single European Copyright Title on the basis of Article 118 TFEU that would apply directly and uniformly across the EU, in accordance with the Commission's objective of better regulation, as a legal means to remedy the lack of harmonisation resulting from Directive 2001/29/EC;

5. Recommends that the EU legislator further lower the barriers to the re-use of public sector information by exempting works produced by the public sector – as part of the political, legal and administrative process – from copyright protection;

6. Calls on the Commission to safeguard public domain works, which are by definition not subject to copyright protection and should therefore be able to be used and re-used without technical or contractual barriers; also calls on the Commission to recognise the freedom of rightholders to voluntarily relinquish their rights and dedicate their works to the public domain;

Exceptions and limitations

11. Calls on the Commission to make mandatory all the exceptions and limitations referred to in Directive 2001/29/EC, to allow equal access to cultural diversity across borders within the internal market and to improve legal certainty;

13. Calls for the adoption of an open norm introducing flexibility in the interpretation of exceptions and limitations in certain special cases that do not conflict with the normal exploitation of the work and do not unreasonably prejudice the legitimate interests of the author or rightholder;

18. Stresses the need to enable automated analytical techniques for text and data (e.g. 'text and data mining') for all purposes, provided that permission to read the work has been acquired;

19. Calls for a broad exception for research and education purposes, which should cover not only educational establishments but any kind of educational or research activity, including non-formal education;

20. Calls for the adoption of a mandatory exception allowing libraries to lend books to the public in digital formats, irrespective of the place of access;

6.2 Open Source

While the language-technology based industry solutions targets an agile high tech industry, it seems that many fields are still dominated by expensive and slowly developing monolithic proprietary software that makes it especially hard for the many SMEs to compete with the 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 systems (e.g., Drupal).

Still, open source projects usually do not run by themselves. They require well thought-through forms of organisation fitting the respective community and type of project. Therefore, these

25 http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-%2F%2FEP%2F%2FNONSGML+COMPARL+PE-546.580+02+DOC+WORD+V0%2F%2FEN

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 want to support the development of the latter for the language industry. In fact, some tools and standards already exist in the indsutry and in language technology research, open source development is the normal case. But 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.

6.3 Language Policy

Technological progress would be even more efficient and effective if the proposed research effort 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 program 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 to achieve funding investment that pays back, ideally PPPs.

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

Numerous US-originating breakthroughs in IT that have subsequently led to commercially successful products of great economic impact could only be achieved by a combination of systematic long-term research support coupled with public procurement. Many types of aircraft or the autonomous land vehicle would not have seen the light of the day without massive military support, even the internet or the speech technology behind Apple Siri heavily benefited 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 greed for originality on the side of the public research funding bodies and their constant trial-and-error search for new themes that might finally help the European IT sector to be in time with their innovations has 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 industrial systems including Google Translate.

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 advocate a combination of

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

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.

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 (see the section on Copyright and IPR).

6.4 Standards and Interoperability

6.5 Skills

7 Organisation of Research

(Includes: hybrid research, agile and continuous development, DevOps)

8 Conclusions – Summary – Next Steps

Appendix

A. References

B. Input Documents

- Philipp Cimiano's presentation "The LIDER Roadmap in a nutshell" at Brussels workshop
- Gerald Cultot's presentation "eHealth services" at Brussels workshop
- Andrew Joscelyne's draft position paper "A Strategic Research and Innovation Agenda for a Conversational European Digital Marketplace" (late 2014)
- Nils Lenke's presentation "Nuance Inc." at an internal DFKI workshop
- Dave Lewis' presentation "Shopping Across the Language Barrier" at Brussels workshop
- Ruben Riestra's report "Multilingual data value chains in the Digital Single Market" on the Brussels workshop
- Alan Mas Soro's presentation "Language Technologies for Europe" at Brussels workshop
- Adomas Svirskas' presentation "Pan-European Electronic Document Platform" at Brussels workshop
- Hans Uszkoreit's presentation "European Platform(s) for Machine Translation and other Language Technologies" at META-NET Platform Strategy Meeting (LREC 2014)
- Xenios Xenophontos' presentation "Online Dispute Resolution Platform" at Brussels workshop
- Sonja Zillner's presentation "cPPP Big Data Value- SRIA" at Brussels workshop
- LIDER Roadmap (current draft version)
- META-NET Strategic Research Agenda for Multilingual Europe 2020
- MLi documents
- QTLaunchPad working paper "European Quality Translation Research 2015: Ongoing Work and Roadmap"
- RockIT documents

C. Detailed Roadmaps

D. Digital Language Extinction in Europe

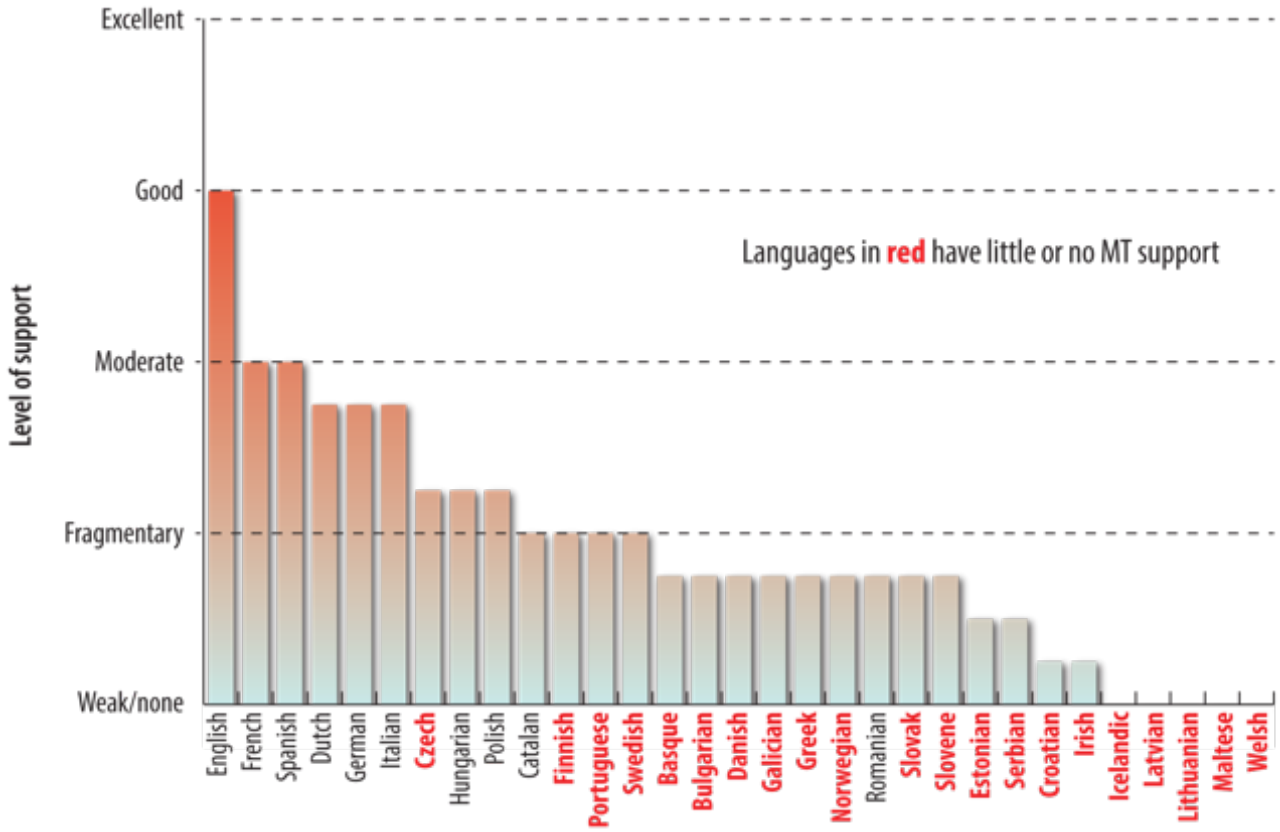
Most European languages are unlikely to survive in the digital age, a study by Europe's leading Language Technology experts warns. Assessing the level of support through language technology for 30 of the more than 60 European languages, we concluded that digital support for 21 of the 30 languages investigated 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, Latvian, Lithuanian and Maltese, receive this lowest score in all four areas. On 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.



E. Key Contributors

F. Milestones and History

5 Presentation: Strategic Agenda for the multilingual Digital Single Market

In the following, we include a presentation given by Georg Rehm at an internal meeting with the EC and LT_Observatory in Luxembourg on March 30, 2015.

CRACKER

Cracking the Language Barrier

META-NET

A META-NET Project

Strategic Agenda for the multilingual Digital Single Market

Georg Rehm

Meeting on Riga Summit and SRIA, Luxembourg – March 30, 2015

This project has received funding from the EU's Horizon 2020 research and innovation programme through the contract CRACKER (grant agreement no.: 645357).
META-NET was formerly co-funded by FP7 and ICT PSP through the contracts T4ME (grant agreement no.: 249119), CESAR (grant agreement no.: 271022), METANET4U (grant agreement no.: 270893) and META-NORD (grant agreement no.: 270899).

Outline

- SRIA: Current State
- Strategic Funding Programme: Overview
- Alternative Proposals for Layer 2
- Next Steps
- Public Consultation Phase

SRIA: CURRENT STATE

MDSM and Data Economy

- Challenge: the DSM doesn't exist yet – it's fragmented into 20+ different language communities and sub-markets.
- LT is a critical enabler for the *multilingual* DSM.
- However, language barriers not yet mentioned in VP Ansip's most recent DSM communications.
- LT as a key contributor to data value chains.
- Big Data Value SRIA: some data sources are multilingual; multilingualism mentioned as barrier for data processing.
- How can LT contribute to the Big Data Value PPP?

SRIA

- Need for an updated vision on how LT can contribute to creating a mDSM and creating effective (multilingual) data value chains.
  - Priority A: Technology solutions for the multilingual DSM
  - Priority B: Data value chains, Big Data, European data economy.
- Community response to very concrete European challenges.
- CRACKER and LT_Observatory drive the SRIA forward.
- Input: Roadmaps, agendas and any other inputs from other initiatives.
- SRIA needs to be aligned with CEF (MT@EC), BDVA.
- SRIA needs to be supported and endorsed by the whole community.

History

- **Nov 11, 2014:** Phone conference with EC; SRIA editorial team: LT_Observatory nominates A. Joscelyne and S. Krauwer; CRACKER nominates H. Uszkoreit and G. Rehm.
- **Dec, 2014 and Jan, 2015:** Discussion about potential SRIA outlines.
- **Early Feb, 2015:** CRACKER suggests to start producing individual text pieces.
- **Early Feb, 2015:** CRACKER starts working on individual text pieces.
- **Feb 08, 2015:** LT_Observatory states they work on Business Solutions.
- **Feb 18, 2015:** Confirmation and agreement between the coordinators of LT_Observatory and CRACKER that it's now time to start the text production phase in both projects concurrently (with regard to the most recent SRIA outline); deadline: early March.
- **Mar 06, 2015:** CRACKER circulates input (SRIA V0.01) to LT_Observatory.
- **Mar 10, 2015:** "Test drive presentation" of SRIA V0.1 and the strategic programme to the LIDER consortium (meeting in Berlin) – very positive feedback; LIDER promises input.
- **Mar 10, 2015:** Agreement between the coordinators of LT_Observatory and CRACKER that we concentrate on *one* document (i.e., the SRIA) and that we should open the public consultation phase soon; LT_O confirms it's no longer working on a "shorter document".
- **Mar 13, 2015:** LT_Observatory provides some remarks and edits on Chapter 2.
- **Mar 19, 2015:** LIDER and FALCON provide substantial input to Sections 4.2 and 5.4.
- **Mar 20, 2015:** LT_Observatory provides proposal for "infrastructures and platforms" layer.
- **Mar 21, 2015:** MULTISENSOR provides some remarks and edits for Section 5.2.
- **Mar 23, 2015:** Foreseen date to circulate SRIA V0.2 (slight delay because of illness).
- **Mar 26, 2015:** LT_Observatory provides input for "infrastructures and platforms" chapter.
- **Mar 26, 2015:** CRACKER circulates SRIA V0.02 to LT_Observatory.

SRIA V0.1

- Version 0.1 (ca. 25 pages) contains:
  - Solid outline
  - Introductory chapter
  - Setup of the overall strategic programme
  - Several key visuals
  - First set of solutions for businesses, public services and societal challenges
  - Some indicative content.
- Proverbial work in progress!
- Next slide: outline of current draft version

Strategic Agenda for the Multilingual Digital Single Market

SRIA Editorial Team

Version 0.1 – March 6, 2015

– internal use only –

This draft document is not yet meant to be circulated widely in the community.

Table of Contents

1 Executive Summary 5
2 The Digital Single Market is a Multilingual Challenge 5
2.1 The Digital Single Market and the European Data Economy 8
2.2 The Economic Power of Language Technology and the Language Industry 9
2.3 A Strategic Programme for the multilingual Digital Single Market 9
2.4 EC and Language Technology – Past and Present 13
2.5 Summary and Conclusions 14
3 Solutions responding to Europe's multilingual Challenges 15
3.1 Technology Solutions for Businesses 15
3.1.1 Unified Customer Experience 15
3.1.2 Cross-Cultural Customer Relationship Management 15
3.1.3 Voice of the Customer 16
3.1.4 Business Intelligence on Big Data 16
3.1.5 Content Curation and Content Production 16
3.1.6 Multimodal User Experience for Connected Devices 16
3.1.7 Smart Multilingual Assistants 16
3.1.8 Ubiquitous Cross-Lingual Communication (BGCtoBGC) 17
3.1.9 Translingual Spaces 17
3.2 Technology Solutions for Public Services 17
3.2.1 Voice of the Citizen – Social Intelligence on Big Data 17
3.2.2 E-Participation 18
3.2.3 E-Government 18
3.2.4 Online Dispute Resolution 18
3.3 Technology Solutions for Societal Challenges 18
3.3.1 Adaptable Interfaces for All 19
3.3.2 E-Health 19
3.3.3 E-Learning 19
4 Enabling Platforms, Infrastructures and Services 19
4.1 Translingual Trusted Cloud Platform for Human and Machine Translation 19
4.1.1 Implementation 19
4.2 European Platforms for Language Technology Services 21
4.2.1 Implementation 21
4.3 Multilingual Meaning and Knowledge Services 21
4.3.1 Implementation 21
5 Research Themes 21
5.1 Research Theme 1: HQ Machine Translation and Human Translation 21
5.2 Research Theme 2: Crosslingual and Multilingual Big Data Text and Speech Analytics 21
5.3 Research Theme 3: Conversational Technologies and Natural Language Interfaces 21
5.4 Research Theme 4: Meaning and Knowledge 21
5.5 Core Resources and Technologies for Language Production and Analysis 21
6 Organisation of Research 21
7 Boosting Innovation 21
8 Horizontal Framework Aspects 21
9 Conclusions – Summary – Next Steps 21
A. References 23
B. Input Documents 23
C. Detailed Roadmaps 23
D. Digital Language Extinction in Europe 23
E. Key Contributors 24
F. Milestones and History 24

SRIA V0.2

- Version 0.2 (53 pages) contains:
  - Improved outline
  - Introductory chapter improved
  - Strategic funding programme polished
  - Several key visuals updated
  - First set of solutions for businesses, public services and societal challenges
  - Much more content than V0.1, some of it indicative.
  - Included input from FALCON, LIDER, MULTISENSOR, LT_Observatory.
- Still the proverbial work in progress!

Strategic Agenda for the Multilingual Digital Single Market

Solutions for Overcoming Language Barriers for a truly integrated European Online Market

SRIA Editorial Team

Version 0.2 – March 26, 2015

– internal use only –

This draft document is not yet meant to be circulated widely in the community.

Table of Contents

1 Executive Summary 5
2 The Digital Single Market is a Multilingual Challenge 5
2.1 The Digital Single Market and the European Data Economy 8
2.2 The Economic Power of Language Technology and the Language Industry 10
2.3 A Strategic Programme for the Multilingual Digital Single Market 10
2.4 EC and Language Technology – Past and Present 15
2.5 Summary and Conclusions 16
3 Solutions responding to Europe's multilingual Challenges 18
3.1 Technology Solutions for Businesses 19
3.1.1 Unified Customer Experience and Cross-Cultural CRM 19
3.1.2 Voice of the Customer 19
3.1.3 Business Intelligence on Big Data 20
3.1.4 Content Curation and Content Production 21
3.1.5 Multimodal User Experience for Connected Devices 21
3.1.6 Smart Multilingual Assistants 22
3.1.7 Translingual Spaces 23
3.1.8 Ubiquitous Cross-Lingual Communication (BGCtoBGC) 24
3.2 Technology Solutions for Public Services 24
3.2.1 Voice of the Citizen – Social Intelligence on Big Data 24
3.2.2 E-Participation 25
3.2.3 E-Government 26
3.2.4 Online Dispute Resolution 26
3.3 Technology Solutions for Societal Challenges 27
3.3.1 Adaptable Interfaces for All 28
3.3.2 E-Health 28
3.3.3 E-Learning 28
4 Enabling Platforms, Infrastructures and Services 28
4.1 Translingual Trusted Cloud Platform for Human and Machine Translation 30
4.1.1 Implementation 30
4.2 Multilingual Meaning and Knowledge Infrastructure 32
4.2.1 Implementation 34
4.3 Natural Language Interaction Services 34
4.4 Text Analytics and Production Services 34
5 Research Themes 35
5.1 Research Theme 1: HQ Machine Translation and Human Translation 36
5.1.1 Novel Research Approaches and Targeted Breakthroughs 36

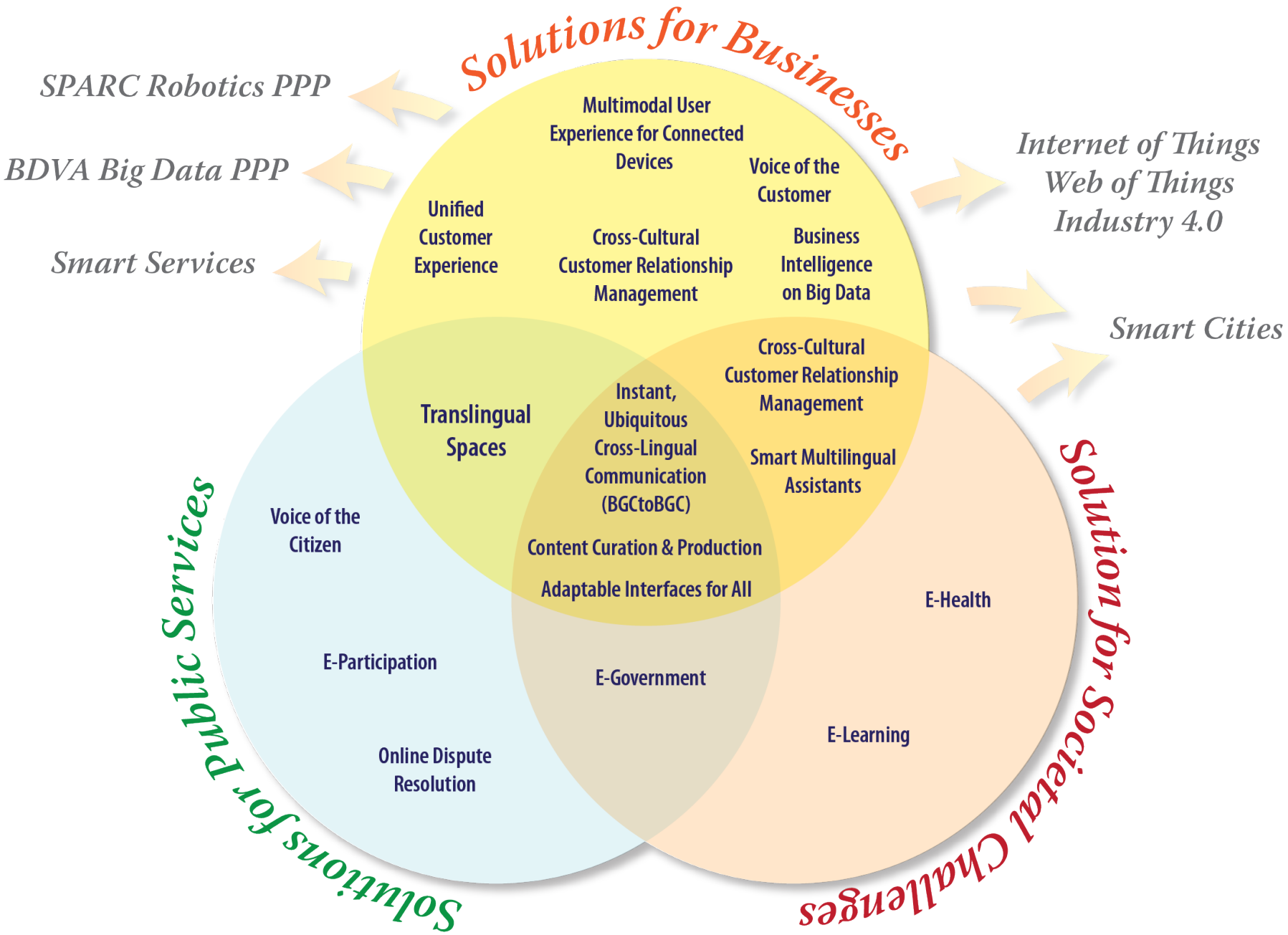
5.1.2 Solution and Realisation 37
5.2 Research Theme 2: Crosslingual and Multilingual Big Data Text and Speech Analytics 39
5.2.1 Novel Research Approaches and Targeted Breakthroughs 40
5.2.2 Solution and Realisation 42
5.3 Research Theme 3: Conversational Technologies and Natural Language Interfaces 43
5.4 Research Theme 4: Meaning and Knowledge 43
5.4.1 Novel Research Approaches and Targeted Breakthroughs 44
5.4.2 Solution and Realisation 45
5.5 Core Resources and Technologies for Language Production and Analysis 46
6 Horizontal Framework Aspects 47
6.1 Copyright and IPR 47
6.2 Open Source 48
6.3 Language Policy 49
6.4 Standards and Interoperability 50
6.5 Skills 50
7 Organisation of Research 50
8 Conclusions – Summary – Next Steps 50
A. References 52
B. Input Documents 52
C. Detailed Roadmaps 52
D. Digital Language Extinction in Europe 52
E. Key Contributors 53
F. Milestones and History 53

Input Documents

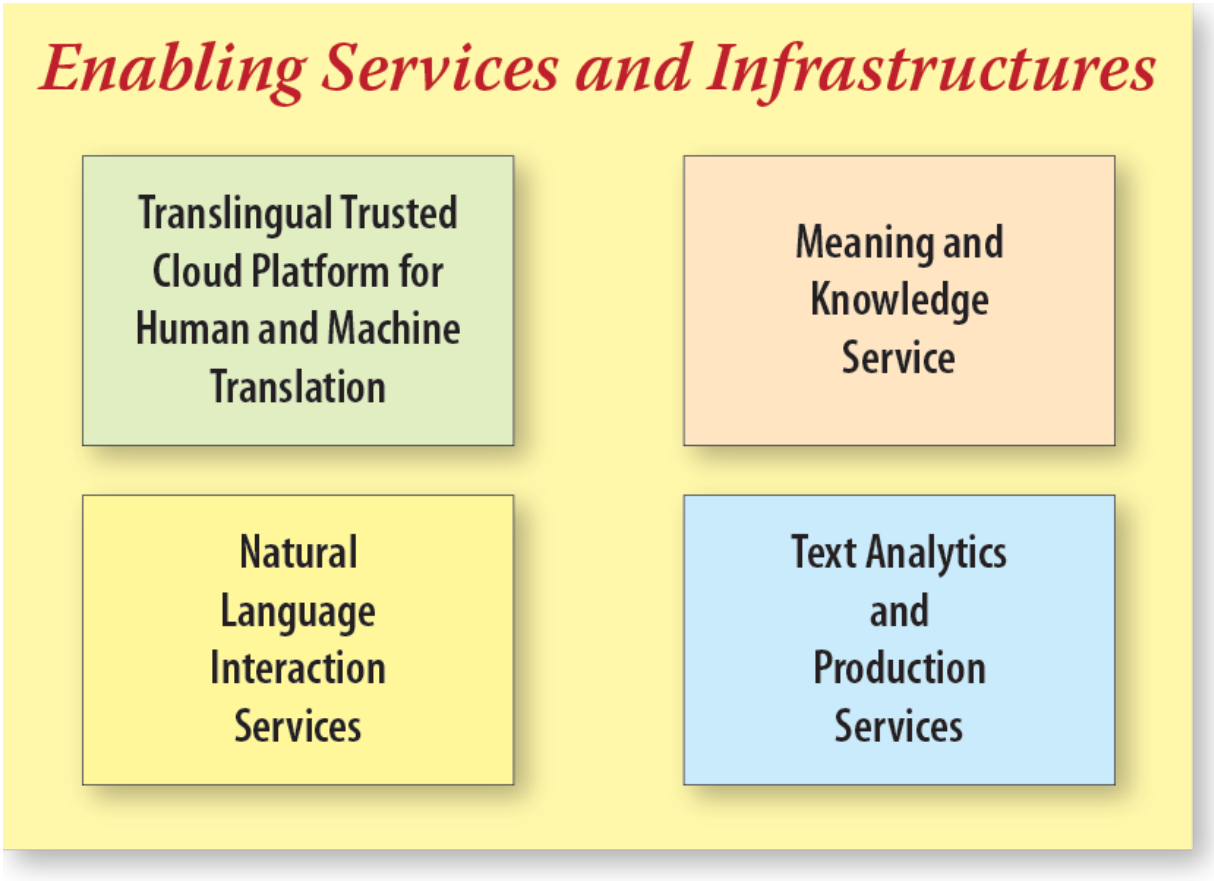
- Philipp Cimiano's presentation "The LIDER Roadmap in a nutshell" at Brussels workshop
- Gerald Cultot's presentation "eHealth services" at Brussels workshop
- Andrew Joscelyne's draft position paper "A Strategic Research and Innovation Agenda for a Conversational European Digital Marketplace" (late 2014)
- Nils Lenke's presentation "Nuance Inc." at an internal DFKI workshop
- Dave Lewis' presentation "Shopping Across the Language Barrier" at Brussels workshop
- Ruben Riestra's report "Multilingual data value chains in the Digital Single Market" on the Brussels workshop
- Alan Mas Soro's presentation "Language Technologies for Europe" at Brussels workshop
- Adomas Svirskas' presentation "Pan-European Electronic Document Platform" at Brussels workshop
- Hans Uszkoreit's presentation "European Platform(s) for Machine Translation and other Language Technologies" at META-NET Platform Strategy Meeting (LREC 2014)
- Xenios Xenophontos' presentation "Online Dispute Resolution Platform" at Brussels workshop
- Sonja Zillner's presentation "cPPP Big Data Value-SRIA" at Brussels workshop
- LIDER Roadmap (current draft version)
- META-NET Strategic Research Agenda for Multilingual Europe 2020
- MLi documents
- QTLaunchPad working paper "European Quality Translation Research 2015: Ongoing Work and Roadmap"
- RockIT documents

STRATEGIC PROGRAMME FOR THE MDSM

Innovative Solutions for the Multilingual Digital Single Market

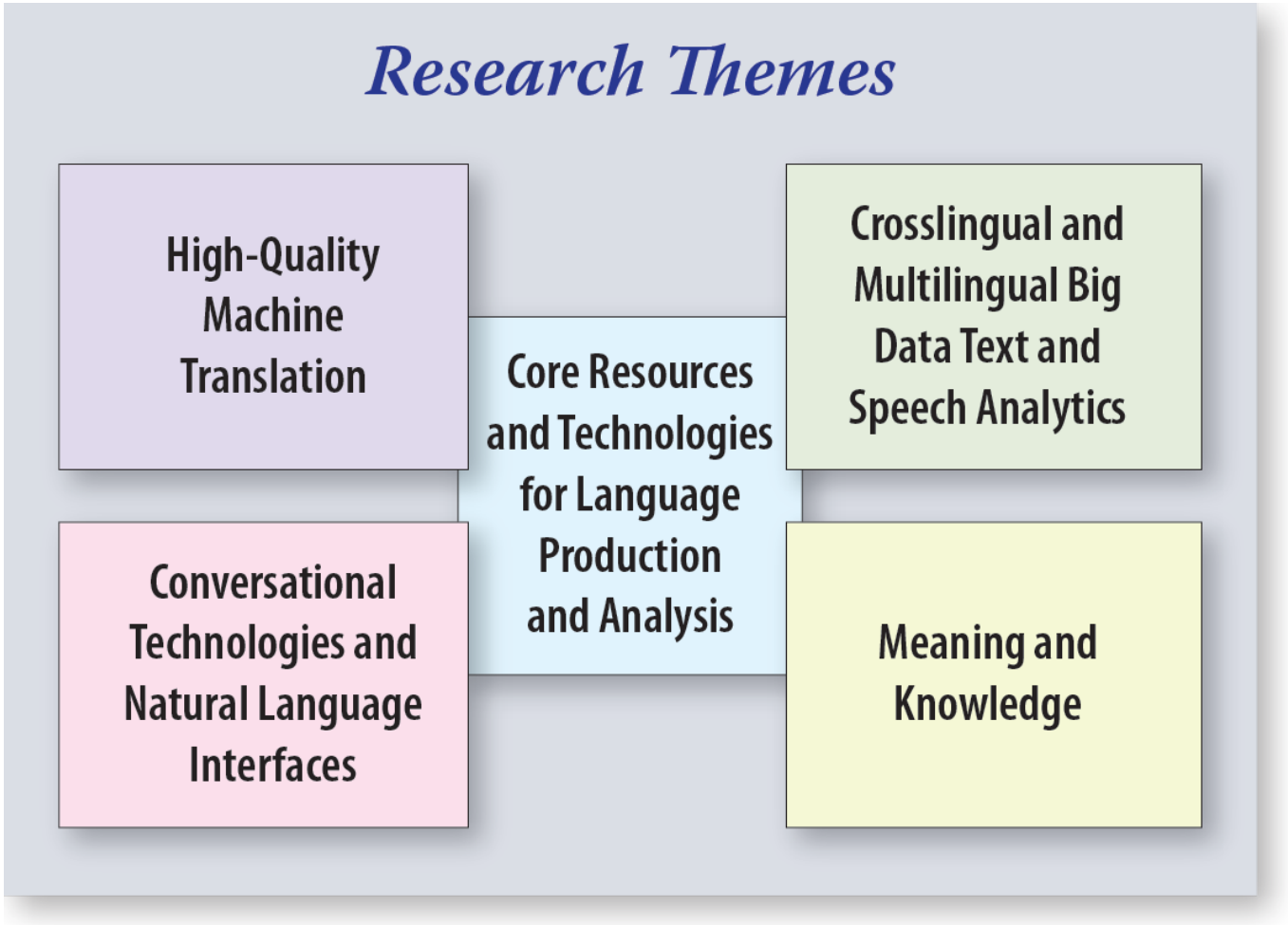


Services and Infrastructures



There is an alternative proposal that will be presented in a few minutes.

Research Themes



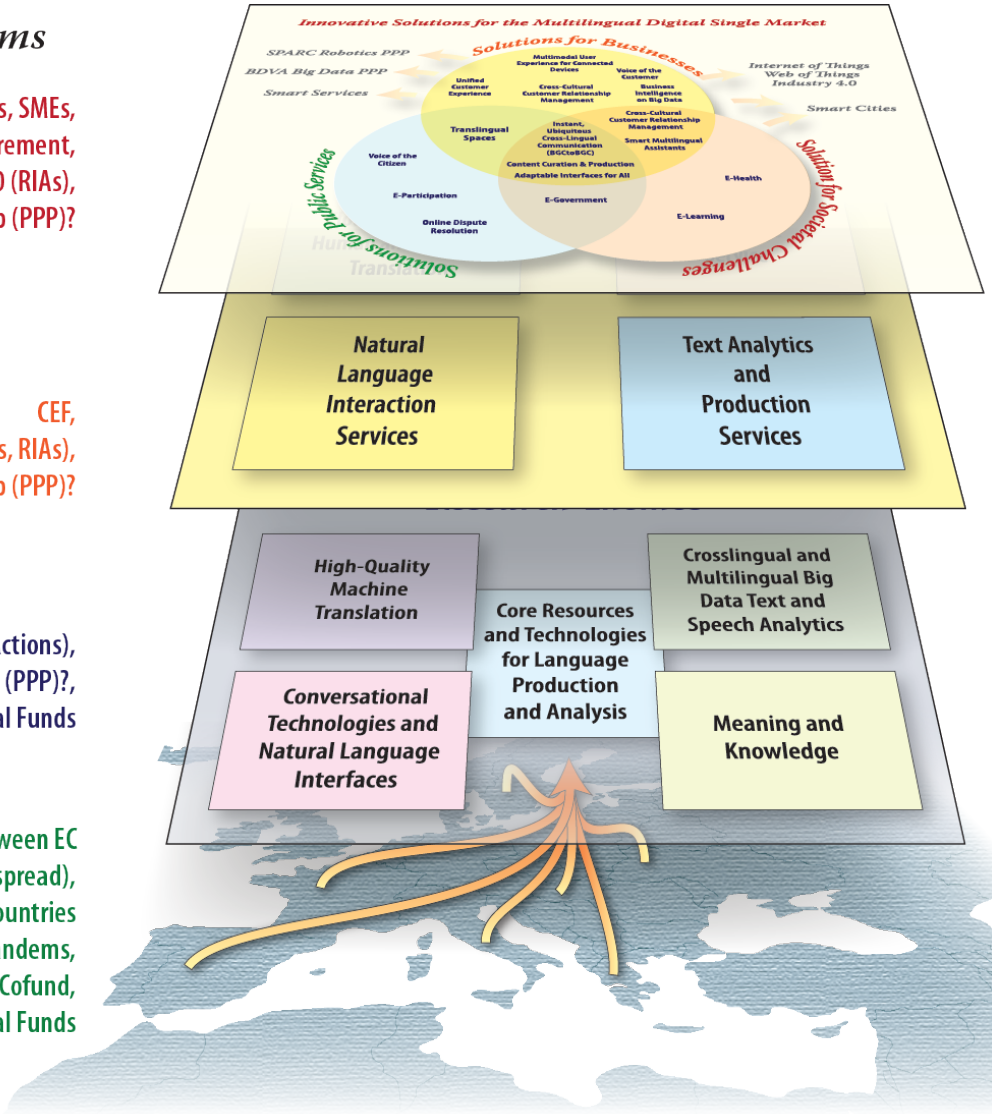
Funding mechanisms

Large enterprises, companies, SMEs, public services and procurement, Horizon 2020 (RIAs), Flagship?, Public-Private-Partnership (PPP)?

CEF, Horizon 2020 (CSAs, RIAs), Flagship?, Public-Private-Partnership (PPP)?

Horizon 2020 (Research Actions), Flagship?, Public-Private-Partnership (PPP)?, EU Structural Funds

Shared funding programme between EC (e.g., through H2020-Widespread), Member States and Associated Countries (technology transfer, research tandems, data procurement), ERA-NET Cofund, EU Structural Funds



Innovative Solutions for the Digital Single Market

enables

Enabling Services and Infrastructures

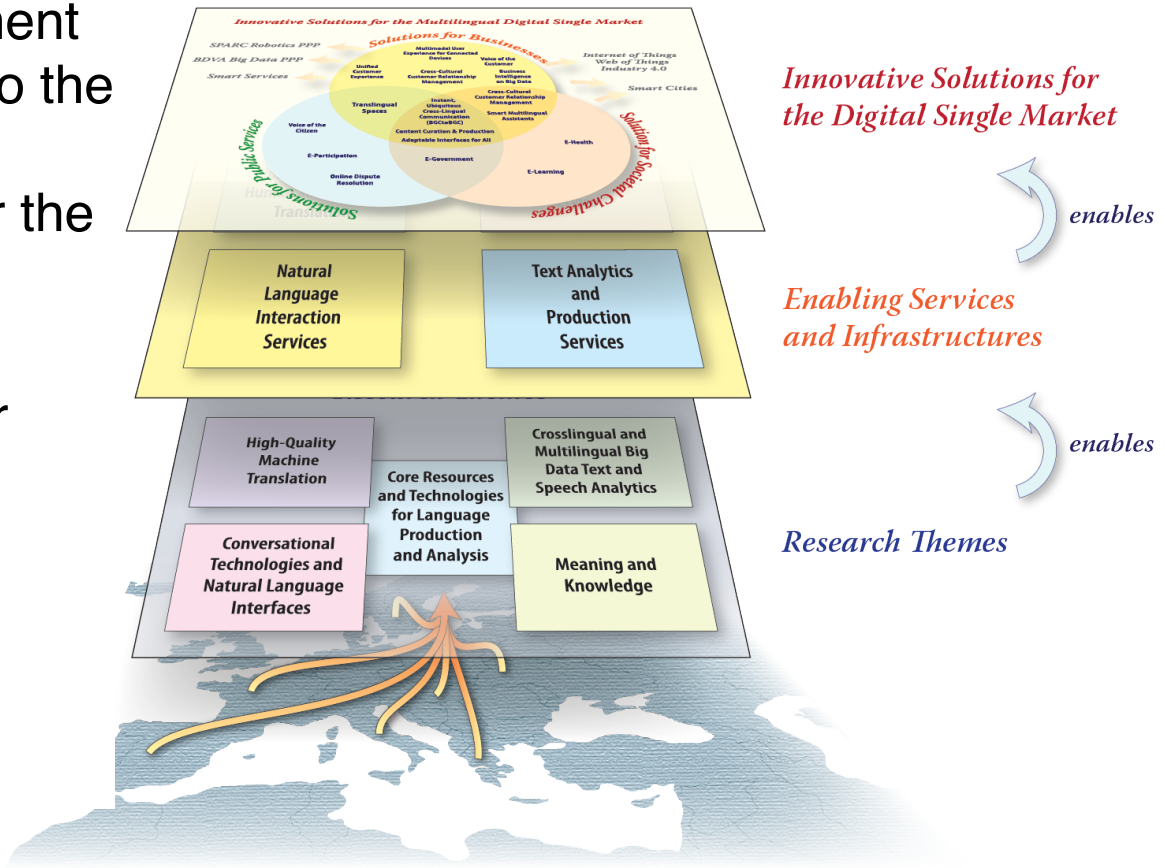
enables

Research Themes

ALTERNATIVE PROPOSAL

LT_Observatory & CRACKER

- Overall a lot of agreement from LT_Observatory to the suggested strategic funding programme for the multilingual DSM.
- LT_O provided an alternative proposal for Layer 2: “Enabling Services and Infrastructures and Platforms”.



“First, Europe needs a basic infrastructure for natural language processing, the **European Language Cloud**. All language processing applications (search, mining, writing, speech, translation, etc.) depend on such basic infrastructure. These are tedious to develop and to maintain, and expensive, since they are required for every single language. The European Language Cloud (ELC) is a public infrastructure which provides the basic functionality required to process unstructured content. Through an API it provides basic language technology services such as tokenization, stemming, part of speech tagging, named entity detection, Identification of measurements, currencies, formulas, etc. for all languages in the same base quality under the same favourable terms.

Second, a **Multilingual Meaning and Knowledge Service** needs to be realised. This service provides seamless and ubiquitous access to multilingual knowledge bases that integrates information about products, companies, places, terms, words, and a plethora of other concepts that are of vital importance for all monolingual, cross-lingual and multilingual language technology components and data value chains. Designing and implementing a general knowledge service is a research challenge but it can become a reality through the combination of existing repositories such as Wikipedia, Wikidata, DBPedia, Linked Open Data sets, WordNet and many other language and data resources. Important for industry and eGov will be sector-specific multilingual knowledge systems which are key assets for serving a global customer base or achieving semantic interoperability.

Third, we need a series of **European Language Technology Application Platforms** for verticals as generic but sector-specific infrastructures. These platforms will provide key industries with a range of language tools and resources that are specifically tailored to the knowledge, linguistic and business process needs of the industry in question. They should be built in coordination with major companies operating in vertical sectors that see an advantage to sharing certain resources with their competitors so as to avoid “reinventing the wheel” and becoming more competitive individually and as a sector. Language technology suppliers who provide the services on these platforms will be able to draw on research and innovation outcomes from other layers to ensure that the technology remains cutting edge. Typical verticals that would be candidates for these services could be automotive, various parts of the healthcare sector, chemicals, legal and financial services, media & publishing, construction.”

Goal of Layer 2 (Infrastructures and Services): Bridge between Research Results and Solutions

- **Technology Solutions for Businesses** (e.g., Smart Multilingual Assistants)
- **Technology Solutions for Public Services** (e.g., Social Intelligence on Big Data)
- **Technology Solutions for Societal Challenges** (e.g., Adaptable Interfaces for All)

- Research Theme 1: **HQ Machine Translation and Human Translation**
- Research Theme 2: **Big Data Text and Speech Analytics**
- Research Theme 3: **Conversational Technologies and Natural Language Interfaces**
- Research Theme 4: **Meaning and Knowledge**
- **Core Resources and Technologies for Language Production and Analysis**

Proposal CRACKER: Infrastructures/Services for Technology Areas

- **Technology Solutions for Businesses** (e.g., Smart Multilingual Assistants)
- **Technology Solutions for Public Services** (e.g, Social Intelligence on Big Data)
- **Technology Solutions for Societal Challenges** (e.g., Adaptable Interfaces for All)

- **Translingual Trusted Cloud Platform for Human and Machine Translation**
- **Text Analytics and Production Services**
- **Natural Language Interaction Services**
- **Meaning and Knowledge Infrastructure**

- Research Theme 1: **HQ Machine Translation and Human Translation**
- Research Theme 2: **Big Data Text and Speech Analytics**
- Research Theme 3: **Conversational Technologies and Natural Language Interfaces**
- Research Theme 4: **Meaning and Knowledge**
- **Core Resources and Technologies for Language Production and Analysis**

Proposal LT_Observatory: Preprocessing, Knowledge, Verticals

- **Technology Solutions for Businesses** (e.g., Smart Multilingual Assistants)
- **Technology Solutions for Public Services** (e.g, Social Intelligence on Big Data)
- **Technology Solutions for Societal Challenges** (e.g., Adaptable Interfaces for All)

- **European Language Technology Application Platforms** (automotive, healthcare, chemicals, legal and financial services, media and publishing, construction etc.)
- **Multilingual Meaning and Knowledge Service**
- **European Language Cloud** (tokenization, stemming, part of speech tagging, etc.)

- Research Theme 1: **HQ Machine Translation and Human Translation**
- Research Theme 2: **Big Data Text and Speech Analytics**
- Research Theme 3: **Conversational Technologies and Natural Language Interfaces**
- Research Theme 4: **Meaning and Knowledge**
- **Core Resources and Technologies for Language Production and Analysis**

Discussion/Questions

Proposal CRACKER

- One stop stop for solutions
- Infrastructures, platforms and services are applicable (base) technologies for all solutions and any verticals (top layer).
- Pre-processing implicit.
- The four research areas (Layer 3) have their own unique platforms and test beds for hybrid research (Layer 2).
- Verticals are left to LT industry (through the Solutions Layer 1).

Proposal LT_Observatory

- Infrastructures share pre-processing.
- Complex services not contained in European Language Cloud (such as, e.g., sentiment analysis)? Are they implemented for each vertical separately?
- No direct connection between research and infrastructures?
- No machine translation?

NEXT STEPS

SRIA: Important Aspects

- SRIA must be endorsed and supported by the whole community.
- This can only be achieved if the whole community can participate in the preparation process.
- Immense pressure and timing constraints.
- Consequence 1: the SRIA must not be prepared by two CSAs only.
- Consequence 2: the public consultation phase must be opened as soon as possible.
- If not, the community will feel excluded and will not support or, worst case scenario, boycott our plans.
- Fortunately, from the few discussions with colleagues and other projects that we had already, the community is eager to help!

Constraints and Expectations

Current State of Play vis-à-vis EC:

- VP Ansip is aware of our work and looks forward to reading the SRIA.
- Commissioner Navracsics (Education, Culture, Youth and Sport) looks forward to reading the SRIA.
- VP Ansip plans to present his DSM strategy in early May. Afterwards it will probably be impossible to introduce new topics in his strategy.

Consequences:

- The first solid version of the SRIA (V0.7) must be ready as soon as possible. Must be attractive, convincing, substantial and meaningful.
- End of April probably too late for SRIA V0.7: we should write a short strategy paper (ca. 3-5 pages, strategic and general) as input for VP Ansip and the DSM team and strategy based on what we know now.

Next Steps (Suggestion)

- Harmonise the two alternative Layer 2 content suggestions (March 30).
- Begin work on short strategy paper for VP Ansip (March 31).
- Finalise first complete draft of SRIA V0.3 (very early April).
- Public SRIA consultation phase (ca. ten days) – collect feedback and get the buy-in and support from the community (early April).
- Further work on the SRIA document (optimising text, graphics, argumentation, including new facts and figures, priorities, timing etc.)
- Include feedback from SRIA public consultation phase (ca. April 15).
- Finalise strategy paper and send it to VP Ansip, DSM team (April 15).
- Finalise and typeset SRIA V0.7 (ca. April 20).
- Present SRIA V0.7 at Riga Summit (April 27-29).
- Public SRIA consultation phase – collect feedback (May).
- Present SRIA V1.0 at Riga Digital Agenda Event (June).

PUBLIC CONSULTATION PHASE

Public Consultation Phase

- Idea: collaborate with the whole LT community (and beyond) on the SRIA online and, ideally, in real time
- Cross-platform, web-based
- Giving specific, contextual feedback with markup tools:
  - Comment: insert text or attach sticky notes to add feedback, ideas, suggestions
  - Annotate: highlight passages, underline key points, or strikeout mistakes
- Invite colleagues to public consultation phase through all channels and mailing lists that we have available.
- Focus on the currently running EU-funded projects.
- Longer pieces of input can be provided/included via email.

Collaboration Tools: Examples

A.nnotate (a.nnotate.com)
Only 30 pages per month free, look and feel not contemporary

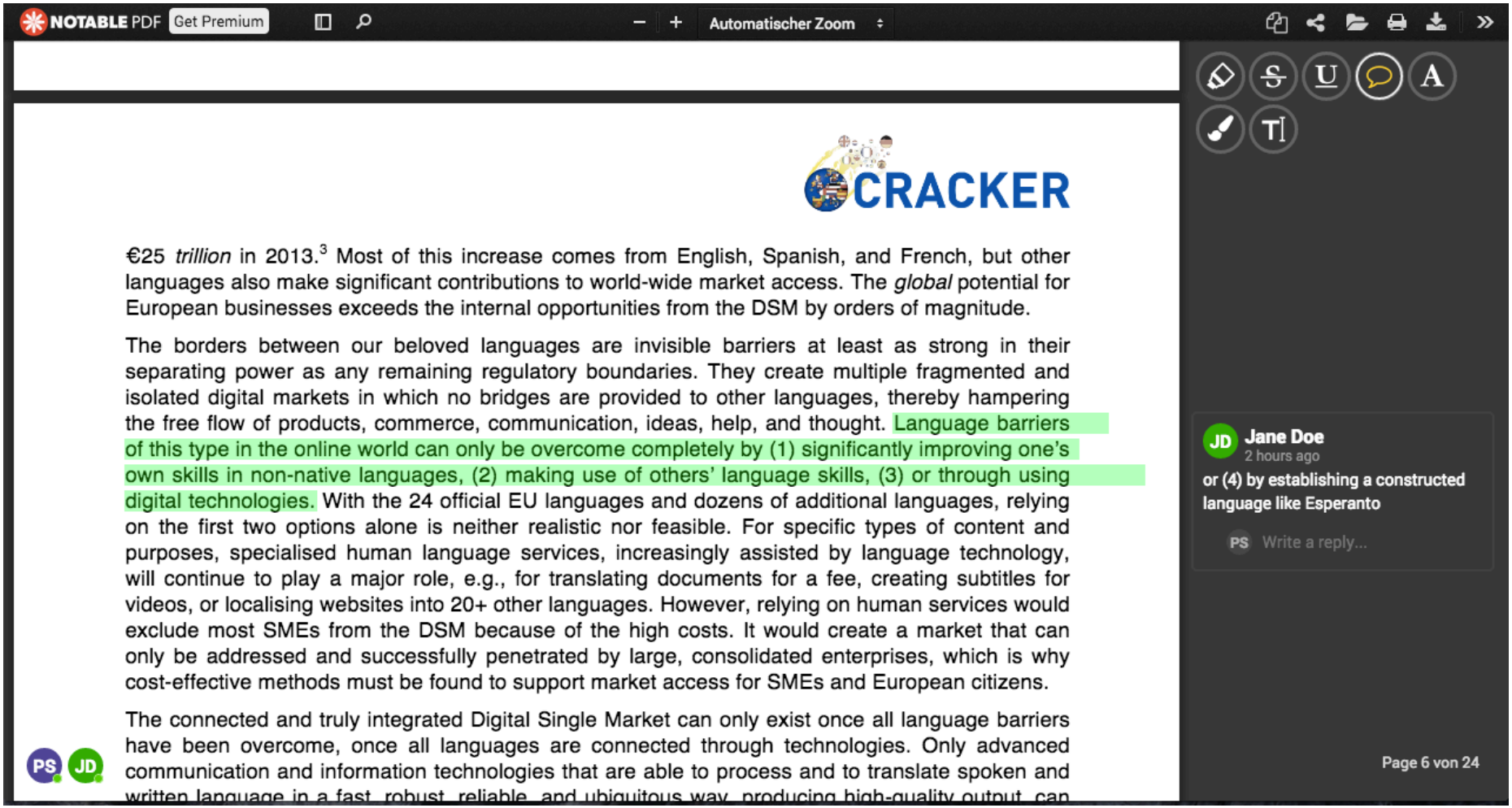
Annotate.co (annotate.co)
Unlimited number of collaborators only in the team version which is very expensive

Dochub (dochub.com)

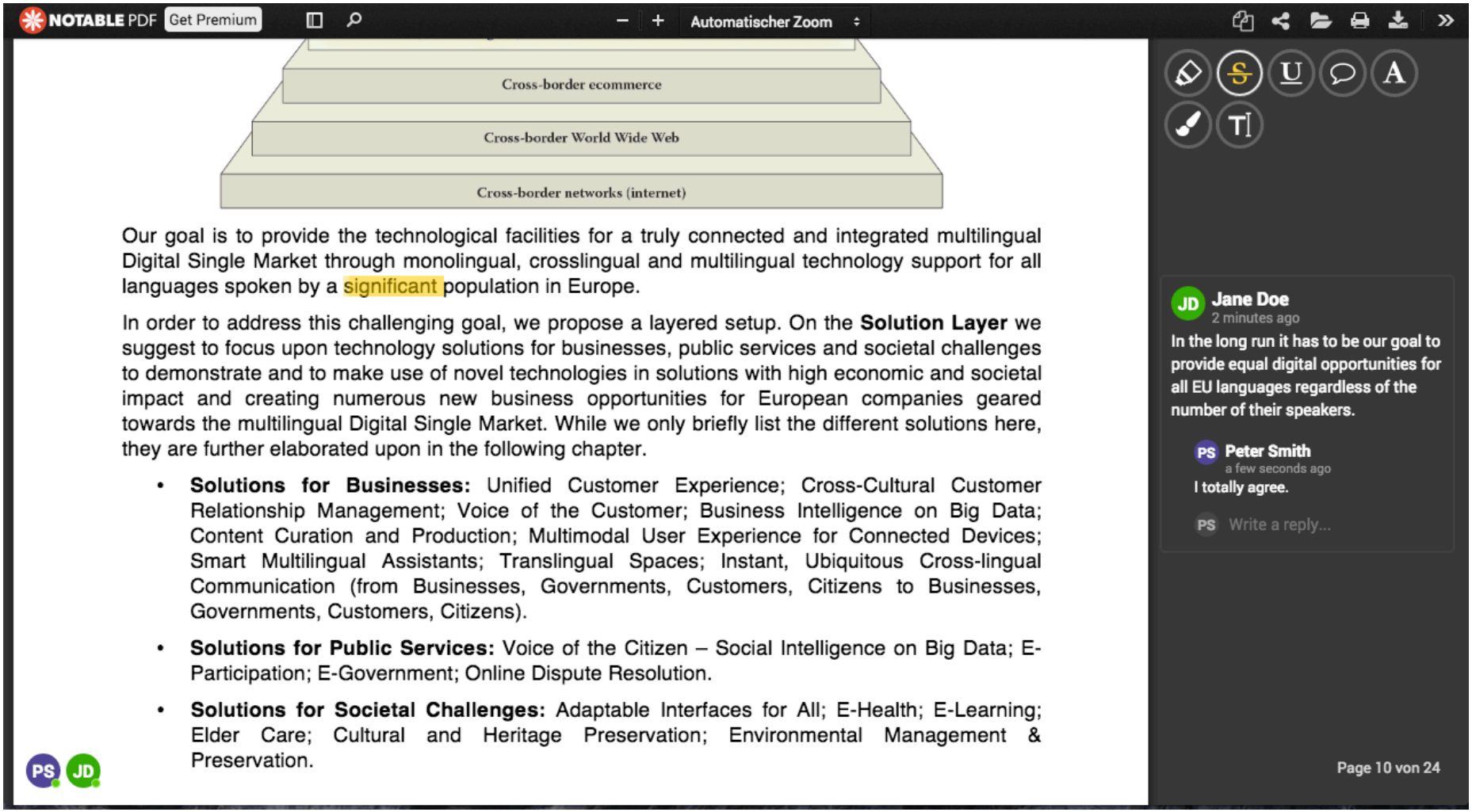
Xodo (xodo.com)
Collaborators have to be invited via email, the public link allows only read-only access

Notable PDF (notablepdf.com) – example: http://bit.ly/19na7aB

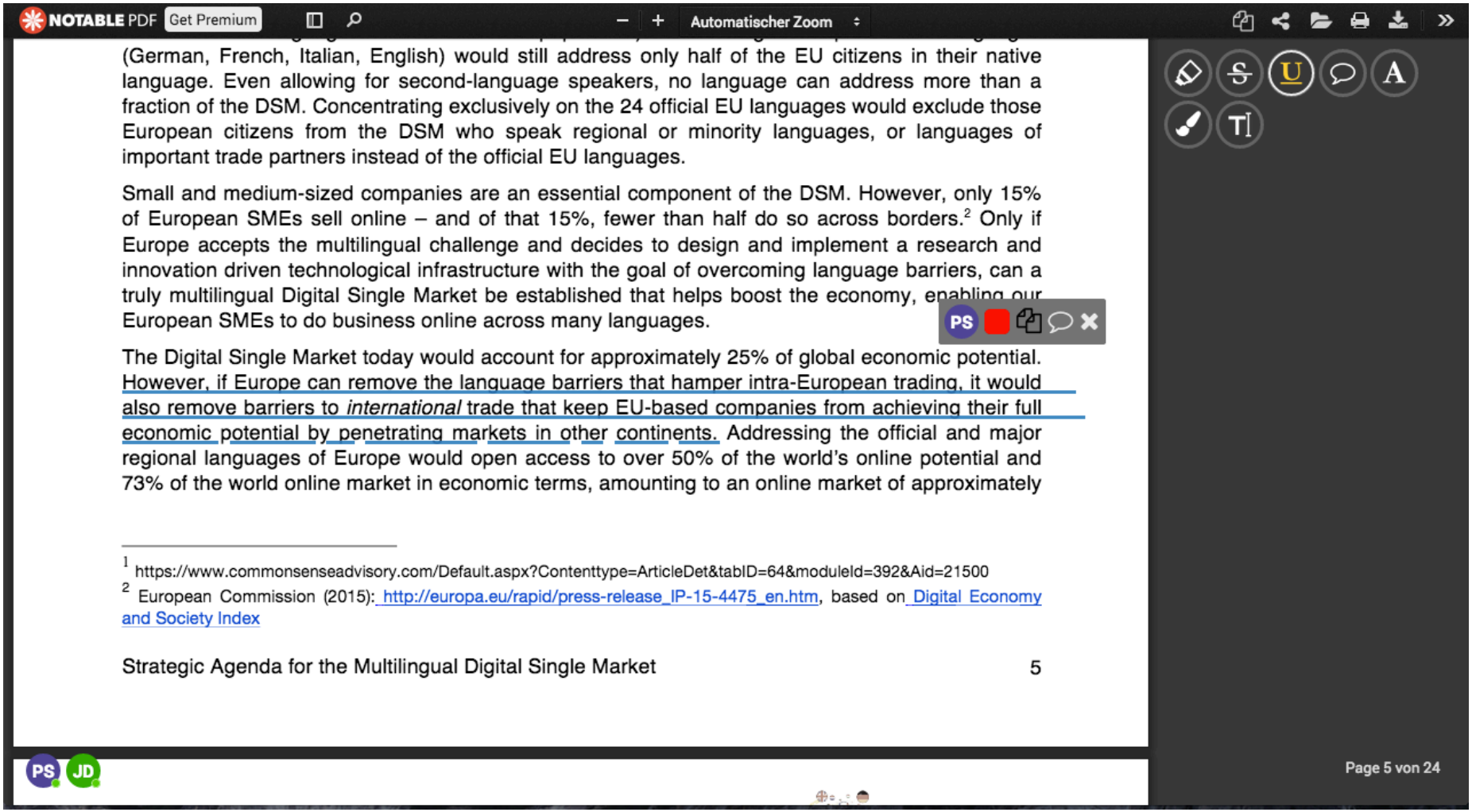
Highlight Text – Add Comments



Reply to Comments



Underline/Comment on Passages



Add Sticky Notes