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 D5.7

Strategic Research and Innovation Agenda for the LT/MT field (final version)

Editor: Georg Rehm (DFKI), Stefanie Hegele (DFKI)

Dissemination Level: Public

Date: 31 December 2017

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	31.12.2017
Actual date of delivery	31.12.2017
Deliverable number	D5.7
Deliverable title	Strategic Research and Innovation Agenda for the LT/MT field (final version)
Type	Report
Status and version	Final; this deliverable reports on version 1.0 of the SRIA
Number of pages	169
Contributing partners	DFKI
WP leader	DFKI
Task leader	DFKI
Authors	Georg Rehm (DFKI), Stefanie Hegele (DFKI)
Internal reviewers	n.a. (SRIA prepared and reviewed in a community process)
EC project officer	Pierre-Paul Sondag (M1-M18), Susan Fraser (M19-M36)
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 Executive Summary of the SRIA (Version 1.0) 5

3 Strategic Research and Innovation Agenda – Language Technologies for Multilingual Europe. Towards a Human Language Project (Version 1.0) 7

4 LREC 2018 Paper: Language Technology for Multilingual Europe: An Analysis of a Large-Scale Survey regarding Challenges, Demands, Gaps and Needs 8

5 Report on the Survey: Language Technology for Multilingual Europe 9

6 Presentation: Language Technologies for Multilingual Europe. Towards a Human Language Project 10

1 Introduction

This deliverable contains the current Version 1.0 of the Strategic Research and Innovation Agenda – "Language Technologies for Multilingual Europe. Towards a Human Language Project". In addition to the SRIA itself (Section 3), the deliverable contains several additional documents: the extended abstract of a paper, which reports on the large-scale CRACKER survey "Language Technology for Multilingual Europe", accepted for publication at the conference LREC 2018 (May 2018, Miyazaki, Japan; Section 4), the full and detailed report on this survey (Section 5), and a presentation that provides more details on the SRIA document, its development, main aspects and next steps (Section 6); this presentation was prepared by the CRACKER Coordinator for META-FORUM 2017 in Brussels.

Building upon past activities, in particular the META-NET Strategic Research Agenda for Multilingual Europe 2020 (published in early 2013) and the LT Innovation Manifesto (June 2014), CRACKER contributed to the elaboration of a Strategic Research and Innovation Agenda for the LT/MT field (SRIA). For the initial SRIA Version 0.5 we collaborated with LT_Observatory. The SRIA defines solution visions and includes economic argumentations on how LT is a key enabler for the Digital Single Market. This preliminary version of the SRIA (Version 0.5) was presented at META-FORUM 2015 and the Riga Summit 2015 event (May 2015).[1] Afterwards we revised the SRIA and presented the results at META-FOURM 2016[2] (Version 0.9) and META-FORUM 2017[3] (Version 1.0). Version 0.9 and 1.0 were prepared, endorsed and published by the Cracking the Language Barrier federation.[4]

At META-FORUM 2017, we presented Version 1.0 of our Strategic Research and Innovation Agenda. The new version formulates novel and modified approaches and solutions in order to make the Digital Single Market multilingual and is titled "Language Technologies for Multilingual Europe. Towards a Human Language Project". The SRIA Version 1.0 was prepared by an editorial team with representatives of the Cracking the Language Barrier federation. The editorial team was chaired by the Coordinator of CRACKER. The SRIA can be downloaded from http://www.cracker-project.eu and from http://www.cracking-the-language-barrier.eu.

The current situation can be summed up as follows. It is very much evident that the language and multilingualism topic is getting more and more visibility and traction in the EC and EP. CRACKER has significantly contributed to this current situation, i.e., to the change towards establishing a firm role for sophisticated Language Technologies and Artificial Intelligence. Now the political will is needed to establish a language policy change on the level of Member States and the EU. The European LT community needs to continue intensifying the push and keeping up the pressure on the Member States, the EP and the EC to initiate the Human Language Project as a concerted action and shared programme between the EU and the Member States.

[1] See Deliverables D1.3, "Report on META-FORUM 2015" and D5.5, "Position Paper and preliminary joint SRIA for the LT/MT field".
[2] See Deliverables D1.4, "Report on META-FORUM 2016" and D5.6, "SRIA for the LT/MT field".
[3] See Deliverable D1.5, "Report on META-FORUM 2017".
[4] See Deliverables D1.2, "Kick-off meeting of the ICT-17 group of funded projects", D1.7, "Final report on coordination among QT projects", D3.7, "Coordination with and support of ICT-17a and ICT-17b projects" and D4.5, "Report on coordination between MT research and CEF".

2 Executive Summary of the SRIA (Version 1.0)

All 24 official EU member state languages are granted equal status by the EU Charter and the Treaty on the European Union. However, omnipresent language barriers still hamper cross-lingual communication and the free flow of knowledge and thought across languages. Multilingualism is one of the key cultural cornerstones of Europe and signifies what it means to be and to feel European. But at the same time Europe is facing multiple challenges. First, its multilingual setup is also one of the main obstacles of a truly connected, language-crossing Digital Single Market as well as Communication and Information Space. The European Language Technology community – including research, development, innovation and other relevant stakeholders – is committed to provide robust and novel technologies in order to successfully turn a fragmented into a truly unified and inclusive Europe, supporting our rich and diverse linguistic heritage. Second, European research in Language Technology is facing increased competition from other continents, especially with regard to recent breakthroughs in Artificial Intelligence. These scientific breakthroughs have lead to commercial successes in the respective regions, which is why many European scientists including young high potentials, are leaving Europe to continue their research abroad.

We recommend setting up the Human Language Project (HLP), a large-scale European Language Technology research, development and innovation flagship programme with the goal of achieving the next scientific breakthroughs for the automatic processing and generation of natural language (both written and spoken). With the rapidly increasing predominance and penetration of Artificial Intelligence in everyday life, the challenge is nothing less but to tackle Deep Natural Language Understanding and Generation by 2030. The HLP is foreseen to be a collaborative endeavour of 10-15 years, coordinated on the European level in close collaboration with the Member States and industry, resulting in numerous service platforms and applications that benefit European society, industry and politics, the Digital Single Market and also European research and innovation. Application areas include: (1) Multilingual E-Commerce, (2) Content, Media, Verticals, (3) Translation, Language, Knowledge, Data. The HLP aims to tightly intertwine basic research, applied research, innovation and commercialisation. Important research themes are (1) Crosslingual Big Data Language Analytics, (2) High-Quality Machine Translation, (3) Meaning, Semantics and Knowledge as well as (4) Conversational Technologies. Public procurement and a policy focus towards “Language Technology-enabled Multilingualism” are crucial and necessary prerequisites for an effective implementation.

The study “Language Equality in the digital age – Towards a Human Language Project” (March 2017), commissioned by the European Parliament, and a recent survey, which represents voices of 634 respondents from 52 countries (including 37 European countries and 27 EU Member States) working on Language Technology, highlight and emphasise the necessity of a HLP tailored specifically to Europe’s needs and demands. With constant political changes posing challenges to a strong Europe and an increased competition from the US and China, it is more important than ever to turn challenges into opportunities.

European Commission VP Andrus Ansip and Director General Roberto Viola (DG Connect) have made several appeals for the need to strengthen multilingualism

through technological innovations. Current EC initiatives, such as the eTranslation building block of Connecting Europe Facility (CEF), and ongoing investment in machine translation do contribute to continuous progress. Language Technology for Europe made in Europe is the key. Not only will it strengthen Europe's place in the pole position of research excellence, but it will contribute to future European cross-border and cross-language communication, economic growth and social stability.

The full SRIA document can be downloaded from http://www.cracker-project.eu and from http://www.cracking-the-language-barrier.eu.

3 Strategic Research and Innovation Agenda – Language Technologies for Multilingual Europe. Towards a Human Language Project (Version 1.0)

In the following we include the final version (Version 1.0) of the document Strategic Research and Innovation Agenda – Language Technologies for Multilingual Europe. Towards a Human Language Project.

Figure 1: Title page of the SRIA Version 1.0

Strategic Research and Innovation Agenda

Language Technologies for Multilingual Europe

Towards a Human Language Project

SRIA Editorial Team

Version 1.0 – December 2017

Cracking the Language Barrier

Cracking the Language Barrier

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

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

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

Organisations

Projects

This document was prepared by the Cracking the Language Barrier federation. It represents the current state of discussion within the language technology research, development and innovation community.

The preparation of this Strategic Research and Innovation Agenda (including previous versions) has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements No. 645357 (CRACKER) and No. 644583 (LT_Observatory).

Executive Summary

All 24 official EU member state languages are granted equal status by the EU Charter and the Treaty on the European Union. However, omnipresent language barriers still hamper cross-lingual communication and the free flow of knowledge and thought across languages. Multilingualism is one of the key cultural cornerstones of Europe and signifies what it means to be and to feel European. But at the same time Europe is facing multiple challenges. First, its multilingual setup is also one of the main obstacles of a truly connected, language-crossing Digital Single Market as well as Communication and Information Space. The European Language Technology community – including research, development, innovation and other relevant stakeholders – is committed to provide robust and novel technologies in order to successfully turn a fragmented into a truly unified and inclusive Europe, supporting our rich and diverse linguistic heritage. Second, European research in Language Technology is facing increased competition from other continents, especially with regard to recent breakthroughs in Artificial Intelligence. These scientific breakthroughs have lead to commercial successes in the respective regions, which is why many European scientists including young high potentials, are leaving Europe to continue their research abroad.

We recommend setting up the Human Language Project (HLP), a large-scale European Language Technology research, development and innovation flagship programme with the goal of achieving the next scientific breakthroughs for the automatic processing and generation of natural language (both written and spoken). With the rapidly increasing predominance and penetration of Artificial Intelligence in everyday life, the challenge is nothing less but to tackle *Deep Natural Language Understanding and Generation by 2030*. The HLP is foreseen to be a collaborative endeavour of 10-15 years, coordinated on the European level in close collaboration with the Member States and industry, resulting in numerous service platforms and applications that benefit European society, industry and politics, the Digital Single Market and also European research and innovation. Application areas include: (1) Multilingual E-Commerce, (2) Content, Media, Verticals, (3) Translation, Language, Knowledge, Data. The HLP aims to tightly intertwine basic research, applied research, innovation and commercialisation. Important research themes are (1) Crosslingual Big Data Language Analytics, (2) High-Quality Machine Translation, (3) Meaning, Semantics and Knowledge as well as (4) Conversational Technologies. Public procurement and a policy focus towards "Language Technology-enabled Multilingualism" are crucial and necessary prerequisites for an effective implementation.

The study "Language Equality in the digital age – Towards a Human Language Project", commissioned by the European Parliament and published in March 2017, and a recent survey, which represents voices of more than 600 respondents from more than 50 countries (including all 27 EU Member States) working on Language Technology, highlight and emphasise the necessity of a HLP tailored specifically to Europe's needs and demands. With constant political changes posing challenges to a strong Europe and an increased competition from the US and China, it is more important than ever to turn challenges into opportunities.

European Commission VP Andrus Ansip and Director General Roberto Viola (DG Connect) have made several appeals for the need to strengthen multilingualism through technological innovations. Current EC initiatives, such as the eTranslation building block of Connecting Europe Facility (CEF), and ongoing investment in machine translation do contribute to continuous progress. *Language Technology for Europe made in Europe* is the key. Not only will it strengthen Europe's place in the pole position of research excellence, but it will contribute to future European cross-border and cross-language communication, economic growth and social stability.

Table of Contents

1. Multilingual Europe – Challenges and Opportunities 1
1.1. Challenge: Digital Single Market 3
1.2. Challenge: Communication across Language Barriers 5
1.3. Challenge: Multilingual Solution for Human Computer/Robot Interaction – Internet of Things and Interactive Voice Interfaces 8
1.4. Challenge: Unprecedented Relevance of Online Media and ICT 8
1.5. Challenge: Making Sense of Big Data 9
1.6. Challenge: Content, Content, Content 10
1.7. EC and Language Technology – Current and recent support 11
1.8. The Economic Power of Language Technology and Services 12

2. Towards a Human Language Project (HLP) 14
2.1. Policy recommendations for Human Language Technology 14
2.2. Overview of the Human Language Project 16
2.3. Economic sectors and application areas of the HLP 17
2.4. Application Examples 17
2.5. Research Focus 18
2.6. Technical Approach 18
2.7. Industry and Research 19
2.8. Timeframe and Costs 20

3. Applications and Solutions 21
3.1. Area: Multilingual E-Commerce 21
3.1.1. Multilingual Application: E-Commerce, CRM and After-Sales 21
3.1.2. Multilingual Application: Online Dispute Resolution 21
3.2. Area: Content, Media, Verticals 22
3.2.1. Multilingual Application: E-Learning 22
3.2.2. Multilingual Application: E-Health 22
3.2.3. Multilingual Application: Content Curation and Content Production 22
3.2.4. Multilingual Application: Written- and Spoken-Language Interfaces 23
3.2.5. Multilingual Application: Voice of the Customer and Voice of the Citizen – Social Intelligence on Big Data 23
3.2.6. Multilingual Application: E-Government 24
3.2.7. Multilingual Application: E-Justice 25
3.3. Area: Translation, Language, Knowledge, Data 25
3.3.1. Multilingual Service: Language Processing, Analysis and Production – Language Resources 25
3.3.2. Multilingual Application: Translation Centre 26
3.3.3. Multilingual Service: Knowledge and Data Repositories 27

4. Research Themes 29
4.1. Research Theme: Cross-lingual Big Data Language Analytics 29
4.1.1. Novel Research Approaches and Targeted Breakthroughs 30
4.1.2. Solution and Realisation 31
4.2. Research Theme: High-Quality Machine Translation 32
4.2.1. Novel Research Approaches and Targeted Breakthroughs 32
4.2.2. Solution and Realisation 33
4.3. Research Theme: Meaning, Semantics, Knowledge 34
4.3.1. Novel Research Approaches and Targeted Breakthroughs 34
4.3.2. Solution and Realisation 35
4.4. Research Theme: Conversational Technologies 36
4.4.1. Novel Research Approaches and Targeted Breakthroughs 36
4.4.2. Solution and Realisation 37

5. Horizontal Topics 38
5.1. Standardisation and Interoperability 38
5.2. Business Models and Ecosystems 38
5.3. Language Policies and Public Procurement 38
5.4. Copyright and Data Protection 39
5.5. Open Source 39
5.6. Related Areas, Applications and Societal Challenges 40

6. Conclusions 41
6.1. Expected Economic Impact 41
6.2. Potential Funding Sources 41

Appendix 43

1. Multilingual Europe – Challenges and Opportunities

Multilingualism is at the heart of the European idea and a true solution to help remove barriers, foster collaboration and create more cultural awareness for a strong and united Europe in its diversity. The recent study "Language Equality in the Digital Age – Towards a Human Language Project"[1], commissioned by the European Parliament's Science and Technology Options Assessment Committee (STOA), recommends, to the European Union, to initiate a new, large-scale European Language Technology research, development and innovation flagship programme, called, in the study, the Human Language Project (HLP).[2]

Based on the results from this study and best practices from previous projects and initiatives such as META-NET[3], we recommend to the whole European Computational Linguistics and Language Technology community to collaborate closely together in order to initiate the Human Language Project (HLP) as a long-term, large-scale, massively funded initiative and unprecedented opportunity for Europe to work on the next generation of Language Technologies. As the key scientific goal it is recommended to strive for full and deep Natural Language Understanding and Generation for all EU official languages by 2030. We foresee setting up a shared programme between the European Commission (crucially, through the framework programme that will succeed Horizon 2020) and the Member States and Regions, and other stakeholders, especially those in industry. The setup needs to include an intertwined mix of basic research, applied research, technology development, innovation and commercialisation; education and talent retention also need to be taken into account. The HLP should run for at least ten years, ideally 15 years, so that the ambitious scientific goal can be adequately addressed. Public procurement and a language-related policy change towards "Language Technology (LT) enabled multilingualism" are crucial related aspects.

The outlined suggestions are further backed by a recently conducted survey titled "Language Technology for Multilingual Europe"[4] , in which the European Language Technology research and innovation community was encouraged to share concrete suggestions and recommendations on how Europe's present challenges can be turned into opportunities in the context of a potential large-scale funded European HLP. Further details and specifics on this will be presented in chapter 2.

The principle that all 24 official languages are supported is perpetuated in the EU Charter (Article 22) as well as in the Treaty on the European Union (art. 3(3) TEU). However, the META-NET White Paper series has revealed that there is a steadily increasing severe threat of digital extinction for a majority of European languages.[5] This deteriorating imbalance between English and the "smaller" languages is exactly where the Language Technology community sees the future's biggest challenge. Raising awareness for the Language Technology potential in Europe on a political level and cultivating the idea of equality has become more important than ever before. This view is also strongly supported by the STOA study. The analysis of 3000 technical, political and strategic documents reveals that the topics of Language Technology and multilingualism are not properly considered in current policies of the EU. Compared to other trending topics Language Technology does not have real significant relevance. The Digital Market Strategy of 2015 only briefly mentions the need of multilingual services.

---

[1] http://www.europarl.europa.eu/RegData/etudes/STUD/2017/598621/EPRS_STU(2017)598621_EN.pdf

[2] While identical in name, the "Human Language Project", as specified in this document, only bears a marginal relationship to previous initiatives bearing the same name. Among others, Abney and Bird (2010) called their "universal corpus of the world's languages" the "Human Language Project". In 2012 TAUS launched their concept for an open language resources and tools platform, which TAUS called "Human Language Project", (see: https://www.taus.net/knowledgebase/index.php/Human_Language_Project). The initiative specified in the STOA report and endorsed by this Strategic Research and Innovation Agenda has a much broader scope and set of objectives than the two initiatives mentioned above.

[3] http://www.meta-net.eu

[4] This large-scale survey project received funding from the EU project CRACKER (grant agreement no.: 645357)

[5] META-NET White Paper Series: http://www.meta-net.eu/whitepapers

In a blogpost called "Multilingualism in the Digital Age: a barrier or an opportunity?"[6] from February 2017, Roberto Viola (Director General DG Connect) and Rytis Martikonis (Director General DG Translation) discuss the current state of play of multilingualism and translation as a public service. Besides commonly used human translation services for legislation, the EU institutions have also continuously invested in machine translation (among other language technologies such as speech recognition and data analytics) in order to provide certain language services faster. The ambitious plan is being operationalised through the programme CEF (Connecting Europe Facility). MT@EC is the Commission's EC system, specifically trained to translate EC documents. It supports a number of European service platforms, e.g. the Online Dispute Resolution platform, the European Open Data Portal and the European e-Justice portal. The next generation of MT@EC, namely eTranslation, is being rolled out and will soon become the automated translation platform of CEF. eTranslation will also become important for social security, eProcurement and eHealth. Given the pace with which technology is advancing, digital solutions (and especially translation technologies) can bridge language barriers, unite a diverse Europe and thereby create a Digital Single Market. The EC Next Generation Internet (NGI) initiative, launched in autumn 2016, responds to the need to represent European social and ethical values in a time where new technologies are rapidly changing everyday life.[7]

We are currently witnessing a highly relevant commercial and industrial interest in Artificial Intelligence, Machine Learning and also Language Technology solutions, especially with regard to technologies based on neural networks. More and more highly interesting and also promising research is, furthermore, carried out in industrial research environments, especially technology enterprises based in the US and in Asia. Many experts in AI perceive cracking human language to be the next barrier and also goal for the next generation of AI technologies. Faced with these challenges and opportunities, the above mentioned survey and STOA study clearly state that in order to stay competitive Europe must get ahead of current developments. Deep learning is a branch of Artificial Intelligence which mimics the activity in layers of neurons in the neocortex, essentially the part of the brain where thinking occurs. A Deep Learning software is trained to recognize patterns in text data, sounds and images. The idea that algorithms can learn this behaviour has been discussed for decades. However, only recent improvements in mathematical formulas, powerful computers and the sheer amount of available data have made new advances possible. Internationally, new breakthroughs in this field are reported on an almost daily basis. In competitive endeavours the big tech companies have already been deploying deep learning methods. Anything powering Google's products and services, like Google Translate, is right now improved by deep learning.[8] Facebook's AI lab has succeeded in developing a question answering system which can handle questions it had never been exposed to. Also, smart speakers like Amazon's Echo make use of deep learning techniques.[9] AI is rapidly taking over many sectors that previously relied on human interaction. Banks are increasingly using chatbots to answer customer queries. For instance, it is suggested that Artificial intelligence will be the main way that banks interact with their customers within the next upcoming years.[10]

All of the above represent an excellent field of opportunity for Europe to become active and to push for even more breakthroughs, paradigm shifts and new, successful solutions and technologies. Europe has a long-standing research, development and innovation tradition with over 800 centres performing excellent, highly visible and internationally recognised research on all European and many non-European languages. For instance, in the field of Machine translation most of the basic research has happened in European research projects. Moses

[6] https://ec.europa.eu/digital-single-market/en/blog/multilingualism-digital-age-barrier-or-opportunity
[7] NGI: https://ec.europa.eu/futurium/en/next-generation-internet
[8] New York Times: https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html?_r=0
[9] Harvard Business Review: https://hbr.org/2017/01/deep-learning-will-radically-change-the-ways-we-interact-with-technology
[10] BBC: http://www.bbc.com/news/technology-39419727

(until 2016 the state of the art phrase-based statistical MT system) and recent Neural Machine Translation (NMT) results of QT21 are just two examples for excellence and world class research.[11] Nonetheless, challenges are omnipresent and must be addressed by the EU, the member States as well as stakeholders from academia and industry.

1.1. Challenge: Digital Single Market

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

The European Commission predicts that the transition to the integrated DSM will deliver up to €400 billion in economic growth by 2020. Measures like eliminating roaming charges, improving legislation (especially copyright and data protection), and making cross-border payments easier are all important and necessary preconditions. However, they are not sufficient to accomplish the overall goal. If customers are hampered by language, online commerce will remain confined to fragmented markets, defined and restricted by language silos. Even the unacceptable suggestion for everyone to use English would not deliver a single market, since less than 50% of the EU's population speaks English, and less than 10% of non-native speakers are proficient enough to use English for online commerce. Approximately 60% of individuals in non-Anglophone countries seldom or never make online purchases from English-language sites; the number willing to purchase from sites in non-native languages other than English is much, much lower.[12]

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

Small and medium-sized European companies are a vital component of the DSM. However, only 15% of European SMEs sell online – and of that 15%, fewer than half, do so across borders.[13] SMEs that sell their products and services internationally exhibit 7% job growth and 26% innovate in their offering – compared to a job growth of 1% and 8% innovation for SMEs that do not sell their products and services internationally.[14] Only if Europe embraces the multilingual challenge and decides to design and to implement research and innovation-driven

[11] http://www.commonsenseadvisory.com

[12] Common Sense Advisory (2014): "Survey of 3,000 Online Shoppers Across 10 Countries Finds that 60% Rarely or Never Buy from English-only Websites".

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

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

technology solutions as well as a service infrastructure with the goal of overcoming language barriers, can the full economic benefits of the DSM be achieved. Enabling and empowering European SMEs to easily use language technologies to grow their business online across many languages is key to boosting their levels of innovation and to help them create jobs.

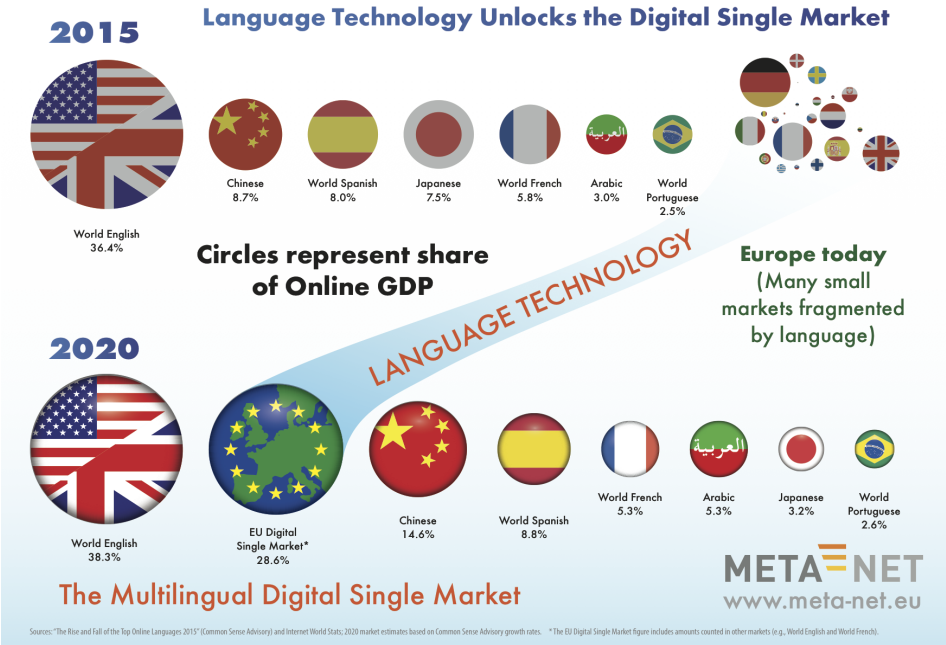


Figure 1: Language technology unlocks the Digital Single Market

The European DSM today would account for approximately 25% of global economic potential. However, if Europe were to overcome the language barriers that hamper intra-European trading, it would also remove barriers to international trade that keep European SMEs from achieving their full economic potential by entering and penetrating markets in other continents beyond our own. Addressing the official and major regional languages of Europe would open access to over 50% of the world's online potential and 73% of the world online market in economic terms, amounting to an online market of approximately €25 trillion (sic) (in 2013).[15] Most of this increase comes from English, Spanish, French, and Portuguese, but other languages also make significant contributions to world-wide market access. The *global potential* for European businesses exceeds the *continent-internal* opportunities from the DSM by orders of magnitude.

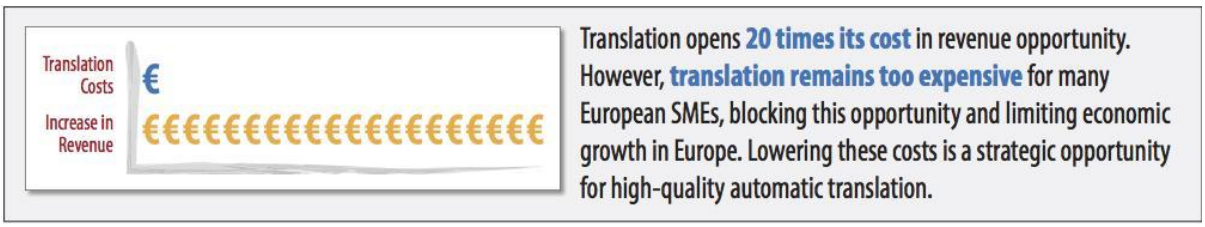


Figure 2: Translation opens 20 times its cost in revenue opportunity

At the end of May 2016, VP Andrus Ansip published a blog post titled "How multilingual is Europe's Digital Single Market?".[16] In his article, VP Ansip not only acknowledges that Europe's multilingualism "brings difficulties for people and businesses to understand each other and to operate across borders". Especially in ecommerce the language barrier can be a concrete obstacle, VP Ansip uses the very adequate phrase "don't understand, won't buy" to describe the situation that especially affects smaller online retailers and web-based traders. As a

[15] Benjamin B. Sargent, Common Sense Advisory (2013): "The 116 Most Economically Active Languages Online", https://www.commonsenseadvisory.com/AbstractView.aspx?ArticleID=5590.
[16] https://ec.europa.eu/commission/2014-2019/ansip/blog/how-multilingual-europes-digital-single-market_en

consequence, online shops often provide (at least) 24 different language versions of their website but it does not end with the actual sale of an item: after-sales services with the same multitude of languages also need to be taken into account. A similar situation exists in the area of data analytics, where data sets in different languages cannot be easily aggregated or semantically processed. VP Ansip mentions that the previous investments of the EC in language technology-related projects (including infrastructural services such as CEF Automated Translation), recent advances in Machine Translation and other multilingual technologies have the potential of breaking through language barriers. Andrus Ansip's goal is "to turn Europe's linguistic diversity from a barrier into an asset" since the DSM "is by definition multilingual".

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

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

1.2. Challenge: Communication across Language Barriers

The borders between our languages are invisible barriers at least as strong in their separating power as any remaining regulatory boundaries. They create fragmented and isolated digital markets with no bridges to other languages, thereby hampering the free flow of products, commerce, communication, ideas, help, and thought. Language barriers in the online world can only be overcome by (1) significantly improving one's own skills in non-native languages, (2) making use of others' language skills, or (3) through digital technologies. With the 24 official EU languages and dozens of additional languages, relying on the first two options alone is neither realistic nor feasible. For specific types of content and purposes, specialised human language services, increasingly assisted by language technology themselves, will continue to play a major role in translating documents, creating subtitles for videos, or localising websites into 20+ other languages. However, relying on human services alone would exclude most SMEs from the DSM because of the high costs involved. It would create a market that can only be successfully penetrated by large, consolidated enterprises, which is why cost-effective methods must be found to support market access for SMEs and European citizens.

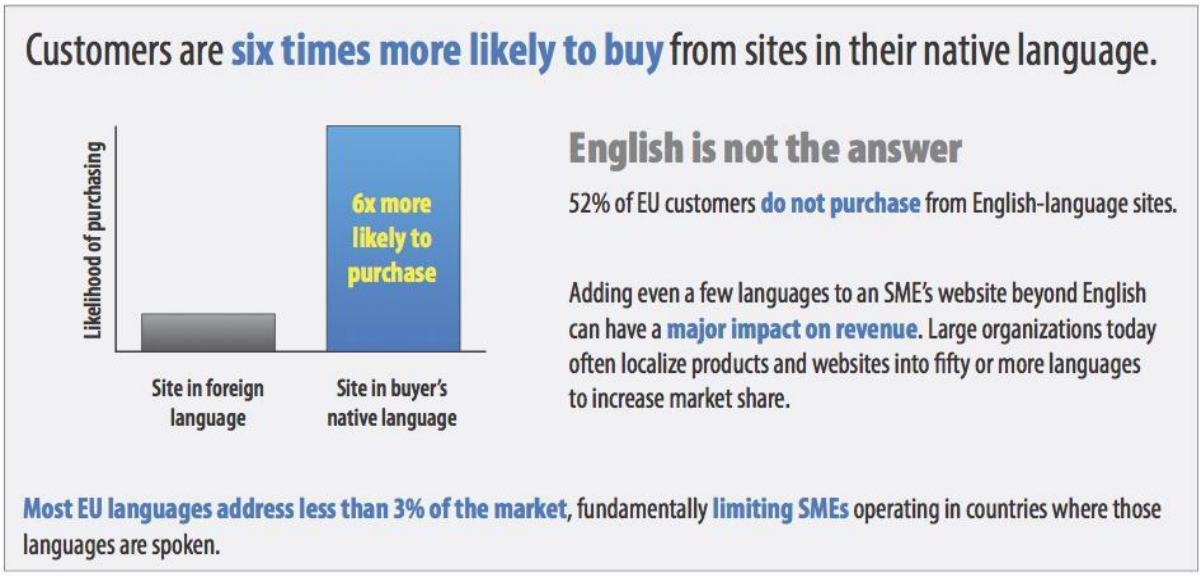


Figure 4: The impact of presenting an online shop in the customer's native language

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

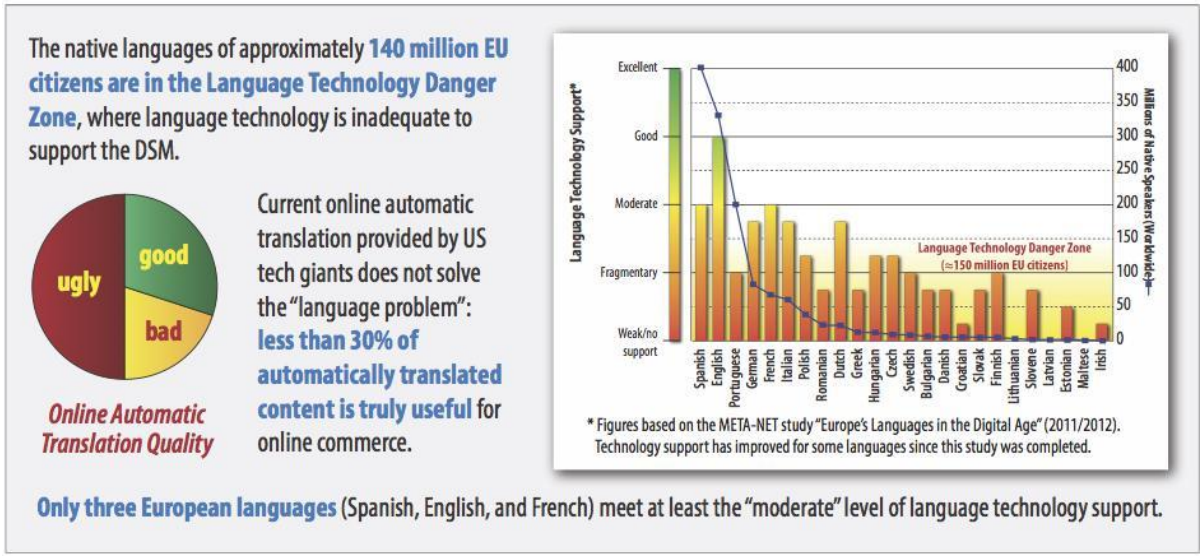


Figure 5: Many European languages only have very weak technology support

A connected and truly integrated Digital Single Market can only exist once all language barriers have been overcome and all languages are connected through technologies. Only advanced communication and information technologies that are able to process and to translate spoken and written language in a fast, robust, reliable, and ubiquitous way, producing high-quality output, can be a viable long-term solution for overcoming language barriers.

Establishing such an infrastructure requires a big collective push that involves designing, implementing and deploying technologies, services and platforms, accelerating innovation, research and efficient technology transfer. While only a few of our languages are in a moderate

to good state with regard to technology support, more than 70% of our languages are seriously under-resourced, actually facing the danger of digital extinction (for example, Maltese, and Lithuanian), even though it must be noted that support for these languages with smaller numbers of speakers is slowly increasing (cf. the Appendix).[17]

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

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

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

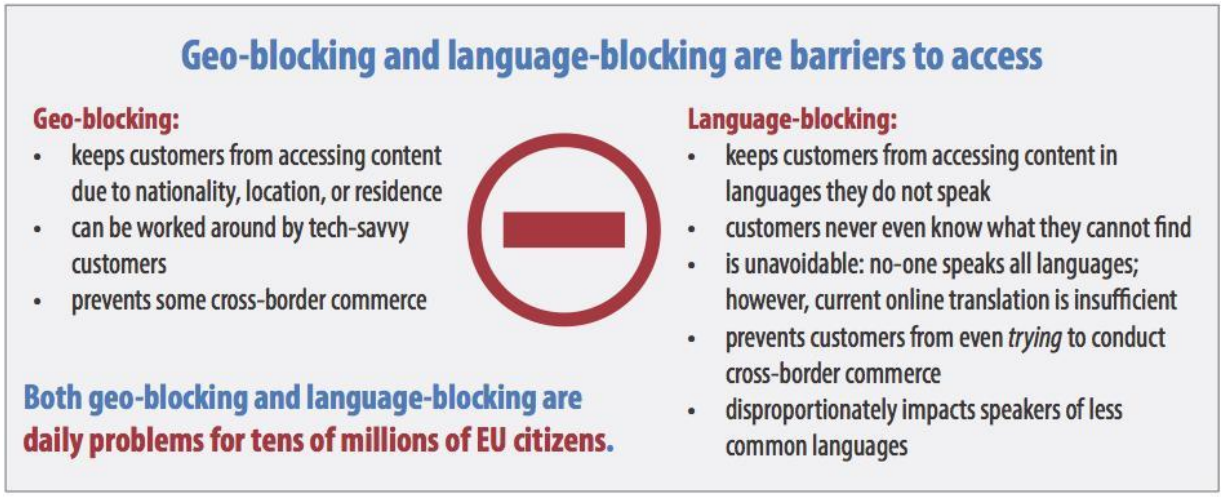


Figure 6: Language-blocking is a barrier to access, just as geo-blocking is

Europe is the most appropriate place for realising the Human Language Project by virtue of applied research, development and 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 working on language, knowledge and interface technologies, and more than 5,000 companies providing language services that can be improved and extended by technology.

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

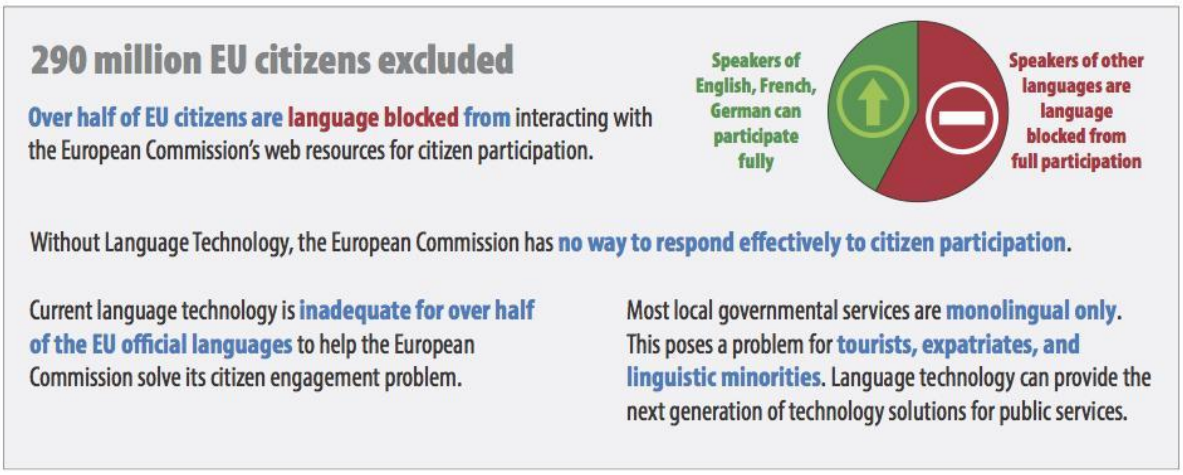


Figure 7: Selected effects of language-blocking

1.3. Challenge: Multilingual Solution for Human Computer/Robot Interaction – Internet of Things and Interactive Voice Interfaces

The Internet of things (IoT) creates smart environments by setting up a network of sensor-driven devices collecting and exchanging data. The European Commission has been committed to supporting the use of IoT technology with a set of policy actions. A recent study by the EC expects the market value of the IoT in the EU to exceed one trillion euros in 2020.

The Digital Market strategy adopted in 2015 emphasises the need to avoid fragmentation and establish interoperability for IoT to unleash its full potential. Policy priorities are based on three pillars: a thriving IoT ecosystem, a human-centred IoT approach and a single market for IoT. A single market will allow citizens to simply plug in and use IoT devices and services anywhere in Europe. In addition to the policy initiatives the EU has set up concrete IoT research and innovation objectives in the ongoing Horizon 2020 programme.[18]

Translating in the future will require an increase in high-speed services that adjust to the ultra-fast Internet of Things (IoT). As translations will be indispensable to connecting 50 billion devices in all languages, language solutions will be included in IoT-linked devices, to assure the flow of communication in various idioms, without interruption and almost instantly.[19]

1.4. Challenge: Unprecedented Relevance of Online Media and ICT

The rise of social media platforms such as Facebook or Twitter has accelerated changes in traditional journalism. Content can easily be published nowadays and the lack of quality checks can mislead or manipulate readers. In a fast-changing world, accurate information is crucial for society. Hate speech, fake news and trolling have become a widespread phenomenon which gained global visibility in the course of the American presidential election in 2016 where, for instance, news was spread about Pope Francis endorsing Donald Trump. The gravity and political impact is being acknowledged by the European institutions even though the steps towards countering the arising challenges are still tentative.[20]

The EU is currently in the process of setting up an expert group dealing with the sensitive issue of fake news. "The Commission needs to look into the challenges that online platforms create for our democracies as regards the spreading of fake information and initiate a reflection on what would be needed at EU level to protect our citizens," Juncker said.[21]

The leveraging of Language Technologies to automatically detect fake news, online propaganda, and hate speech has received growing interest. While current state-of-the-art

[18] https://ec.europa.eu/digital-single-market/en/policies/internet-things
[19] https://blog.soprasteria.com/internet-of-things-multilingual-communication
[20] http://www.europarl.europa.eu/RegData/etudes/ATAG/2017/599384/EPRS_ATA%282017%29599384_EN.pdf
[21] https://www.euractiv.com/section/digital/news/gabriel-to-start-eu-expert-group-on-fake-news/

NLP and AI methods can handle tasks such as machine translation, summarisation etc. at a fairly high level, the challenge of identifying claims and making judgements about their validity and accuracy presents a truly multilingual challenge.

1.5. Challenge: Making Sense of Big Data

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

Figure 8: Multilingual data value chains

We also need to pay attention to the sheer volume of data generated. Only one hour of customer transaction data at Wal-Mart, corresponding to 2.5 petabytes of data, is 167 times the amount of data housed by the Library of Congress.[22] Data growth keeps rising: 90% of the data available today has been generated in the past two years only.[23] 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.[24] The Internet of Things will not only add more but also additional types of data (including large amounts of textual data, of course): Cisco estimates that currently less than 1% of physical objects are connected to computer networks. According to recent estimates by Cisco this number will rise to up to 50 billion connected devices by 2020, corresponding to between 6 and 7 devices per person on the planet. Europe needs a scalable technological infrastructure for handling its big

[22] Beñat Bilbao-Osorio et al. (ed.) (2014): "The Global Information Technology Report 2014 – Rewards and Risks of Big Data", World Economic Forum and INSEAD.
[23] SINTEF (2013): "Big Data, for better or worse: 90% of world's data generated over last two years".
[24] John Gantz and David Reinsel (2012): "The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East", International Data Corporation (IDC).

data sets. While the specific Big Data solutions circling around computer science and advanced database technologies will be taken care of by the Big Data Value Contractual Private Partnership (BDV cPPP), the examples given above demonstrate the need for complementary language technologies and for creating synergies between the BDV cPPP and the European Language Technology community by including robust and precise multilingual text analytics technologies that can perform at web-scale level and, even more crucial, at an Internet of Things level.

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

1.6. Challenge: Content, Content, Content

There is increasing pressure to overcome language barriers: online content in hitherto dominant languages is declining and "long-tail" languages are rising.[25] In line with the constant rise of content, absolute numbers are rising for all languages, and much more significantly so for less spoken languages. For example, Basque, Galician, and Catalan all have an increasing share vis-a-vis Spanish; even though the numbers are small, they indicate a long-term shift.

This trend goes hand in hand with increasing public demand for content in regional or local languages due to the increasing availability of broadband, as well as mobile connectivity and rising numbers of online users and online services. Europe's citizens are no longer satisfied with using only a few major languages. As a consequence, businesses that cannot provide content in local languages will be global losers. Market saturation for dominant languages has been reached, any additional growth is coming from outside the established markets, historically served by a smaller set of languages.[26] If we extrapolate the trends reported by Common Sense Advisory, it only took 37 languages to reach 98% of the world online population in 2009, but already 48 in 2012. The predicted number in 2015 was 62 languages.

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

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

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

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

while in France. Here the impact of language barriers is measured not just in terms of Euros but in terms of health and, potentially, lives.

1.7. EC and Language Technology – Current and recent support

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

In recent years, the EU has been supporting projects such as EuroMatrix, EuroMatrixPlus (2006-2008, 2009-2012), Let'sMT! (2010-2012), and iTranslate4 (2010-2012), which draw on basic and applied research along with industrial collaboration to generate MT resources for many European languages. More recently, the large-scale META-NET initiative (supported in its first phase by four EU projects), which started in 2010, has assembled the LT community around its core network of excellence which consists of 60 research centres in 34 European countries: META, the Multilingual Europe Technology Alliance, has more than 800 members. META-NET has prepared studies such as its 30-volume White Paper Series, and the META-NET Strategic Research Agenda for Multilingual Europe.[27] The open resource exchange infrastructure META-SHARE provides access to thousands of language resources and technologies. The EU has also facilitated the coalescing of the LT industry through the FP7 support action LT COMPASS. The resulting industry association, LT-Innovate, currently counts 180 corporate members. LT-Innovate issued a Report on the State of the European Language Technology Industry[28] and an Innovation Agenda[29]. At the beginning of 2015 new projects have been launched, funded through the Horizon 2020-ICT 17 call. In addition to the large research action QT21, which is working on new paradigms for high-quality machine translation, three innovation actions are adapting and applying new MT methods for industrial and commercial use cases. In the middle of 2015, the EU project CRACKER initiated the "Cracking the Language Barrier" federation of organisations and projects working on technologies for a Multilingual Europe. This umbrella initiative is continuously getting more members and currently assembles 10 organisations and more than 20 projects.

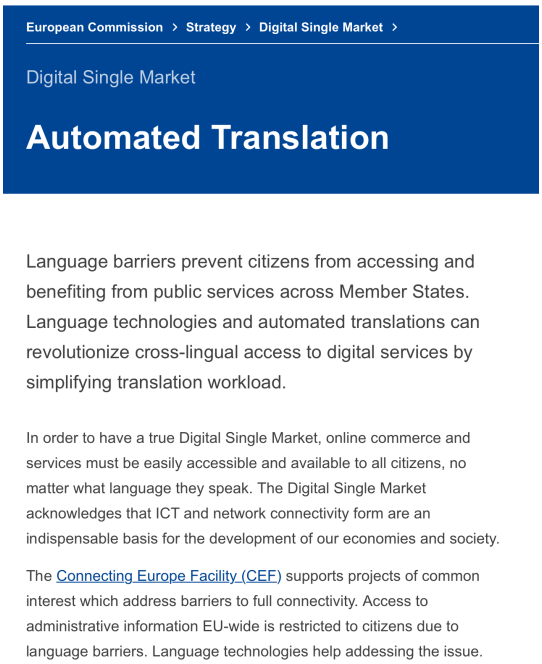
In parallel to the research and innovation-oriented activities funded through FP7 and Horizon 2020, the EC is further advancing the Connecting Europe Facility programme (CEF). Part of CEF Telecom is the Automated Translation building block that "helps European and national public administrations exchange information across language barriers in the EU" (see excerpt from blog entry by DG Connect in Figure 10 below) and also to make all of CEF's Digital Service Infrastructures multilingual.[30] This Automated Translation service, CEF AT, builds on an existing MT system, MT@EC, developed within the EC (DG Translate). It is being implemented on the Moses toolkit, under the Interoperability Solutions for European Public Administrations (ISA) programme. One of the key ideas is to harness the linguistic knowledge embodied in the EC's database of translated documents covering the 24 official languages of the EU. MT@EC is currently only available to staff members of the EC and the EP as well as public administrations of EU member states. A closer collaboration between CEF AT and the European LT community has been established through the service contract ELRC (European Language Resource Coordination) which was awarded in September 2015, especially with regard to the systematic and coordinated collection and exploitation of language resources in all CEF participating countries. A follow-up project has been approved until 2019.

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

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

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

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


European Commission > Strategy > Digital Single Market >

Digital Single Market

Automated Translation

Language barriers prevent citizens from accessing and benefiting from public services across Member States. Language technologies and automated translations can revolutionize cross-lingual access to digital services by simplifying translation workload.

In order to have a true Digital Single Market, online commerce and services must be easily accessible and available to all citizens, no matter what language they speak. The Digital Single Market acknowledges that ICT and network connectivity form are an indispensable basis for the development of our economies and society.

The Connecting Europe Facility (CEF) supports projects of common interest which address barriers to full connectivity. Access to administrative information EU-wide is restricted to citizens due to language barriers. Language technologies help addessing the issue.

Figure 10: Blog post by DG Connect Team, last updated on 9 May 2017

Looking beyond the EC, research by TAUS[31] has shown that European research funding that fostered the development of the open source MT toolkit Moses has opened up new business opportunities in language technology by enabling companies to reduce the cost required to translate content, particularly in fields such as technical support. These cost reductions have helped companies to increase their multilingual reach and engage with customers in language markets inaccessible through traditional translation routes. There is a clear long-term trend to increasing language support and increasing customer engagement via language technologies. According to the report, there are already 22 operative Moses-based MT companies with an estimated market share of about $45 million or about 20% of the entire MT solutions market.

The immediate future holds huge prospects for Language Technology funding through the new H2020 WP 2018-2020 calls. While the European Commission had, unfortunately, dropped Language Technology as a dedicated topic for WP 2016/2017, approximately 12-15 of the upcoming projects do feature some degree of a language component. 25 million Euros will be available for Language Technology projects in call ICT-29-2018, “A multilingual Next Generation Internet”.[32] The European Commission is looking for proposals regarding the topics European Language Grid (one Innovation Action) and domain- specific/challenge-oriented Human Language Technology (several Research and Innovation Actions). The proposed Innovation Action for a European Language Grid would be in line and meet precisely the needs for better open source and infrastructure services requested by the community. The main action point is the creation of an architecture for an open and interoperable grid (ensuring appropriate handling of legal and organisational issues). While the existence of call ICT-29-2018 is a step in the right direction, the available budget does not reflect the commercial and industrial interest in the topic. A related promising development is that the topics of other WP 2018-20 calls are also highly relevant for our field, especially under the umbrella of current AI methods and technologies.

1.8. The Economic Power of Language Technology and Services

In addition to being a key enabling technology for the multilingual DSM, Language Technology comes with a non-trivial economic power itself. The European market for translation,

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

[32] http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/ict-29-2018.html

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 LT industry[33] was evaluated at €26.5b in 2015, projected to rise to €65b by 2020. The size of the overall global language industry in 2016 is estimated at $40 Billion (USD), with estimates of up to $45 Billion by 2020.[34]

Driven through interactive dialogue systems and smart personal assistants such as Apple Siri, Microsoft Cortana, Amazon Echo and Google Home, the global speech technology market alone will reach ca. US$20.9 billion by 2015 and ca. US$31.3 billion by 2017. Yet, this existing capacity is not enough to satisfy current and future needs, e.g., with regard to translation. Today, Google Translate translates the same volume per day as all human translators on the planet translate in one year and is used by more than 200 million people every month.[35] In the last 10 years, Google Translate has grown from supporting just a few languages to 103. Google Translate is now used by over 500 million people, with 100 billion words translated daily.[36]

The setup of the European Language Technology field is very different to the U.S. model. No LT company in Europe is even of comparable size to the big players like Google or Baidu. The EU has to step in to establish similar task, topic and technology coverage to what US, Chinese, Asian companies cover, freely available to SMEs in the EU. Increased public investment is needed to set incentives for young people and early-stage researchers to stay in Europe. In order to make Europe a more attractive work place traditional teaching methods should be enhanced by more relevant research-based experience which also enables EU-wide networking, taking advantage of existing European networks such as CLARIN[37]. Equally important is to increase the size and improve the quality of available language resources by giving continuous support for management, preservation and evolution. In addition, shared investment into HPC and university computing facilities is needed as well as new hardware to support Deep Learning.

A future in which AI will play a crucial role is also challenging for policy making as it is difficult to predict economic and social effects at this point. The White House published a report in early 2017 on “Artificial Intelligence, Automation, and the Economy”[38] which asses the expected impact of AI and its benefits. Experts predict among other things an aggregated productivity growth, the need for different job market skills as well as the loss labour intensive jobs, an uneven distribution of impact across sectors and a general transformation of the job market and creation of new roles. In order to tackle these challenges three broad strategies are suggested for the U.S. economy: the first one is to increase investment in AI, the second focuses on educating people with new skills and the third is to sustain the empowerment of workers ensuring shared growth.

One of the main objectives of the above mentioned STOA study was also to advice on policy suggestions related to Europe’s multilingual challenge and current advances in AI. The study warns that language barriers are likely to have a number of significant social and economic consequences: (1) Foster a language divide, (2) Hamper worker’s mobility, (3) Hindering the access to cross-border public services, (4) Limiting citizens’ engagement and participation in the political process, (5) Creating fragmented markets for cross-border trade and e-commerce, particularly for SMEs.

[33] Figures from LT2013: Status and Potential of the European Language Technology Markets, April 2013
[34] https://www.gala-global.org/industry/industry-facts-and-data
[35] http://googleblog.blogspot.de/2012/04/breaking-down-language-barriersix-years.html
[36] https://www.blog.google/products/translate/ten-years-of-google-translate/
[37] https://www.clarin.eu
[38] https://obamawhitehouse.archives.gov/sites/whitehouse.gov/files/documents/Artificial-Intelligence-Automation-Economy.PDF

2. Towards a Human Language Project (HLP)

Our treasured multilingualism, one of the main cultural cornerstones of Europe and what it means to be and to feel European, is also one of the main obstacles of a truly connected Multilingual Digital Single Market. With language technologies, there are realistic opportunities to remove hampering language barriers while preserving cultural and linguistic diversity and creating a truly integrated EU. The European Language Technology community – including research, development, innovation and other relevant stakeholders – is committed to provide the technologies to achieve this goal.

According to the STOA study no single policy can effectively draw upon Human Language Technology (HLT) and the challenges of Europe's multilingual set up described in the previous chapter. Therefore, the study presents 11 recommendations (R1-R11) which can be structured into five main policy groups: (1) Institutional policies, (2) Research policies, (3) Industry policies, (4) Market policies and (5) Public Service policies.

2.1. Policy recommendations for Human Language Technology

The following briefly summarises these recommendations. More information including the evidence and arguments for the recommendations and assessment criteria for the feasibility and effectivity of their realization can be found in the STOA study.

The main recommendation, the idea to launch a multidisciplinary European Human Language Project, will be further elaborated in the course of this chapter.

Institutional policies

(R1) New institutional framework for HLT

The first recommendation encourages a dedicated newly-established unit within the European Commission for all topics related to "Multilingualism and Language Technologies". Under the responsibility of a EU Commissioner, this new unit should be the umbrella organisation for all policies regarding HLT and thereby provide all European public and private bodies with advisory services.

(R2) Create tools to properly evaluate HLT policies

Further, it is recommended to closely monitor and assess all newly introduced policies. This should be done with a set of appropriate measures agreed on beforehand. Methods to evaluate the policy effects over a longer period of time could entail the analysis of feedback collected through specific surveys and the creation of open datasets containing information that reveal how prominent language barriers still are.

Research policies

(R3) Refocus and strengthen research in the Human Language Project

This recommendation supports the setting up of a large-scale Human Language Project and the accompanying goal of achieving Deep Natural Language Understanding and Generation by 2030. It envisions close collaboration of basic research, applied research, development, innovation and commercialisation for ground-breaking methods, paradigms and approaches as well as technologies and products. (More details regarding research themes, stakeholder involvement and funding will be provided later.)

(R4) European HLP Platform of data and services

It is proposed to draw upon existing infrastructures and platforms such as META-SHARE and CLARIN in order to foster the development of additional resources and to successfully create an open cloud-based platform supporting all European languages. The regulation of the use of such data needs to be made more open and standardised formats should be introduced for core language resources.

(R5) Bridge the technology gap between European languages

The last recommendation of the research policies group focuses on creating true language equality which should be sustained by increased collaboration between linguistic communities in terms of research and technology transfer.

Industry policies

(R6) Foster and support the development of investment instruments and accelerator programs targeting HLT start-ups

This recommendation responds to the lack of coordinated public-private support when it comes to funding for smaller languages and for concrete HLT product development. Specific accelerator programs targeting HLT start-ups are needed in order to preserve the cultural heritage of all languages. With a truly unified Digital Single Market private investors would also be more inclined to invest in technologies for smaller languages.

(R7) Increase the availability of qualified personnel on HLT

Recommendations to prevent the ominous shortage of technical professionals circle around ideas for new incentives aiming to make work in Europe more attractive for researchers. Some of the suggested incentives include the facilitated processes for founding start-ups, a refocus in education relevant for HLT and a general increase of awareness about HLT, especially when it comes to social implications.

Market policies

(R8) Raise awareness of the benefits of HLT

As for market policies, it is recommended to increase efforts to raise awareness on how HLT can help overcome language barriers. Better HLT would not only benefit public bodies and policy makers, but also SMEs which could benefit from access to a pan-European market and an integrated DSM. Citizens could vastly profit from available multilingual public services.

(R9) Promote the automated translation of e-commerce web sites

In order to empower SMEs it is recommended to seize opportunities to increase market size in the wholesale and retail sector by providing all service in multiple languages. A suggestion is to provide economic incentives to SMEs, to use automated translation for their e-shops and developers on the other hand to develop cloud-based services allowing a smooth integration of HLT in E-commerce applications.

Public Service policies

(R10) Public Procurement of Innovative Technology and Pre-commercial Public Procurement

Regarding Public Service policies it is suggested to follow Public Procurement of Innovation (PPI) methods, meaning that the public sector acts as early adopter and purchases innovative solutions not yet available on large-scale commercial basis. In addition, Pre-commercial Procurement (PCP) is an R&D tool that applies when innovative goods and services are not yet available in the market. This procurement process implies that the buyer (in this case the public administration) and the supplier share risks and benefits under market conditions. In this way the public sector can also steer the development of new solutions directly towards its needs.

(R11) Automated translation of national and regional public websites across Europe

It is highly recommended that public services at all levels (local, national and European) use HLT and provide information, websites and digital services in all European languages.

2.2. Overview of the Human Language Project

The HLP will be guided by a comprehensive roadmap which foresees close collaboration between basic research, applied research and innovation. Key contributions come predominantly from the fields of Computational Linguistics, Natural Language Processing, Artificial Intelligence, Language Technology, Linguistics and Computer Science. Another important aspect of this large-scale initiative is the shared coordination and responsibility between the European Parliament, the European Commission and the Member States. Goals of the HLP are the development of new, ground-breaking methods, paradigms and approaches as well as the fostering of technologies, products, innovation, the economy and education (see Figure 11 below).

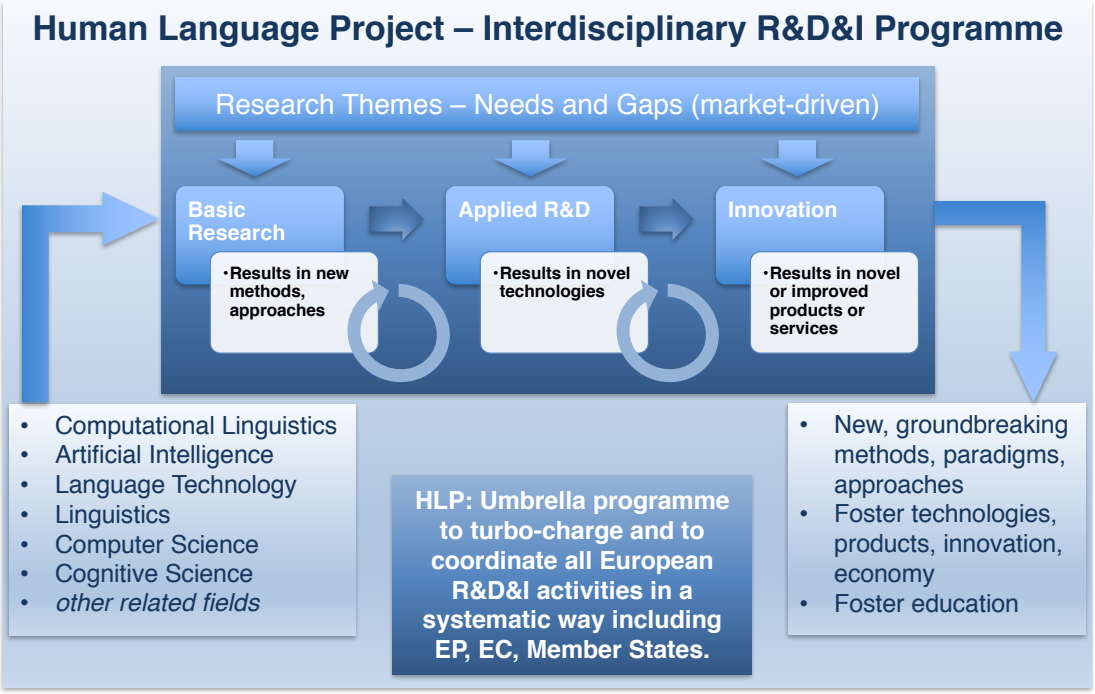


Figure 11: Human Language Project (Rehm, 2017)

The outline of the HLP as presented here in this chapter is informed and inspired by the input and feedback collected from the European LT community through a large-scale survey on Language Technology for Multilingual Europe conducted in June 2017. The survey created a total of 634 responses with a wide demographic reach from 52 countries. The idea of a HLP received substantial support from the group of respondents with 97% stating that they are in favour of establishing such a programme. Further key quantitative and qualitative findings from the survey are summarised in Figure 12 below.

According to the survey, around 16% of the respondents see the biggest challenge that the European LT community is currently facing being the neglect of smaller languages. This is a severe threat, which is leading to a fragmented rather than a united and multilingual Europe. Around 90% state that they work with English in their research (not exclusively though) since they are often given little incentive to solely focus on smaller or minority languages. For instance, when it comes to publishing research results there is a strong bias towards incorporating results for English. Other challenges include the insufficient amount of data resources (approx. 13%), an unwillingness of collaboration within the community (approx. 8%) and, as already indicated above, a lack of funding (approx. 8%).

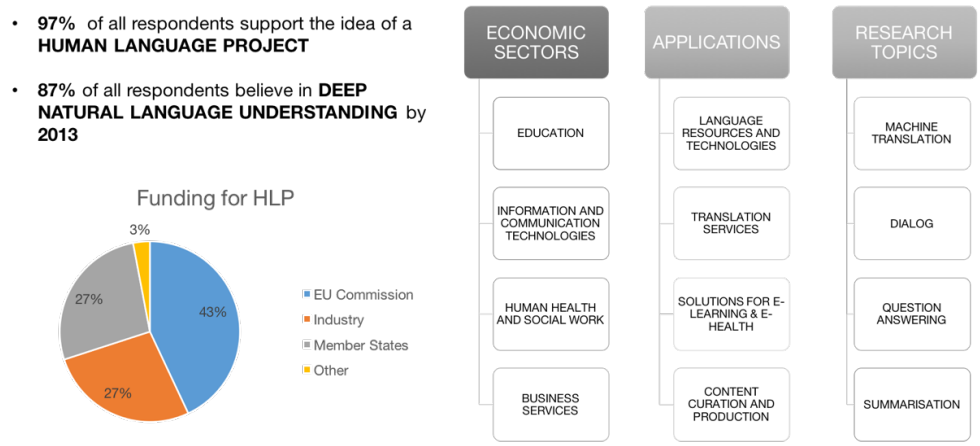


Figure 12: Overview of "Language Technology for Multilingual Europe" survey results

2.3. Economic sectors and application areas of the HLP

Economic sectors where the survey respondents see the highest potential for LT applications, opportunities for commercial growth or promising target markets are Education (71%), Information & Communication Technologies (64%) as well as Human Health & Social Work (45%). Specific services and applications that could benefit the Multilingual Digital Single Market comprise better Language Resources and Technologies (73%), Translation Services (46%), Multilingual Solutions for E-Learning (41%) and E-Health (38%). In the context of industries, sectors and verticals the necessity of an on-going knowledge transfer and effective collaboration between academia and industry is highlighted. The Health sector is unequivocally the most significant one, Education comes in second, closely followed by Tourism and Travel as well as Law and Justice.

Thus, the Human Language Project should focus on building a foundation for the three main Application areas that relate to the three main pillars of the Multilingual Digital Single Market.

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

2.4. Application Examples

Services are needed that provide flexible multilingual technologies – even including *human* translation. These services need to be designed from the outset with special emphasis on high-quality output, trust, data security, reliability, privacy, data protection and confidentiality. We also need a bridge to the world of knowledge, data and meaning through corresponding services. This bridge needs to provide seamless and ubiquitous access to multilingual knowledge bases that integrate information about products, companies, places, terms, words, and a plethora of other concepts that are important for all monolingual, cross-lingual and multilingual language technology components and data value chains. The design and implementation of such a knowledge service is a challenge but it can become a reality through the combination of repositories such as Wikipedia, Waikato, BabelNet, DBPedia, Linked Open

Data sets and other resources and data sets. Important for industry and e-government will be sector-specific multilingual knowledge systems, which are an essential prerequisite for serving a global customer base. Another candidate for one of the sets of generic services is concerned with sophisticated methods for text analytics and production, such as, for example, report generation, text classification, sentiment analysis and opinion mining. As a third example, we foresee conversational technologies and natural language interaction services for dialogue systems and interactive voice interfaces that include the analysis and synthesis of spoken language. All these services would need to be linked up with one another. The spoken language services, for example, would need to contain bridges to the translation services. All services would need to have a 24/7 availability and provide web-scale performance and connectivity to carry out their main purpose: support and enable the application solutions and to support research and innovation by testing and showcasing results as well as providing an environment for hybrid research and devops, i.e., integrating operational services and research.

2.5. Research Focus

The main priority of the HLP is to tackle Deep Natural Language Understanding and Generation by 2030. A significant number of survey respondents (87%) believe this to be a feasible challenge. Research themes will be discussed in further detail in chapter 4. In terms of concrete suggestions for basic and applied research and innovation the following ideas and visions were brought forward.

As far as basic research is concerned a majority mentioned the further development of existing resources (incl. corpora, ontologies, dictionaries etc.) and improvement of data annotations (approx. 9%). In this context, effective legal frameworks for better accessibility are also necessary. Besides, basic research should be centred around deep learning and neural networks (approx. 7%) as well as Natural Language Understanding (approx. 7%). A majority also highlighted the need to further work on existing NLP tasks and tools such as Question Answering, Summarisation, Information Extraction and Sentiment Analysis (approx. 6%).

Applied research should strongly focus on MT according to around 13% of all respondents. Seen as crucial is thereby, again, the improvement of multilingual resources, data sets and terminology repositories, allowing for standardisation and interoperability (approx. 10%). In addition, there is a demand for improved open-source platforms with a wide range of available systems and applications and truly open and unencumbered data and code repositories (approx. 4%).

When it comes to innovation the inclusion of all languages and fostering of intercultural systems is regarded as a top priority (9%). This also presupposes better and stronger relations between academia and industry (7%). Also stressed is the need to bring together knowledge and methods developed for different fields and domains, e. g., e-health, e- government and e-justice (5%). In addition, there is an interest for more advanced visualisations and interfaces, new innovative tools incorporating NLU and seamless human- computer as well as human-robot interactions (5%).

2.6. Technical Approach

The Human Language Project will result in a set of services (in the sense of cloud-services, web-services, RESTful services, application programming interfaces etc.) that drive the Multilingual Applications. They can be conceptualised as Software-as-a-Service, but also as components that can be integrated into stand-alone software. When it comes to strategic guidance what can be derived from the survey responses is the strong suggestion to concentrate funding on smaller scale projects, starting bottom- up with smaller goals, and also to avoid heavy bureaucracy. Hence, we suggest to start with a small set of clearly defined, mission-critical services needed by the majority of applications. This initial set of seed services, then, needs to be able to scale organically into one or more bigger platforms. It will be important

to provide flexibility through a highly innovative ecosystem that enables the emergence of more complex sets of services and platforms (including free and commercial services).

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

When it comes to stakeholder engagement the trend goes towards involving all stakeholders, namely data providers, LT providers and LT consumers. *Users of the technologies* will be able to receive customised integrated services without having to install, combine, support and maintain any software. They will have access to specialised solutions even if they do not use these regularly. *Language technology providers* will have ample opportunity to offer stand-alone or integrated services through component technologies or cloud-based APIs. *Researchers* will have a virtual laboratory for testing, combining, and benchmarking their technologies and for exposing them in realistic trials to real tasks and users. Through the involvement of *users*, valuable data will be collected within these inherently European platforms (vs. Platforms that physically reside on other continents) that can directly feed back into improved services. *Providers of services* that can be enabled or enhanced by text and speech processing will utilise the platform for testing the needed LT functionalities and for integrating them into their own solutions. *Corporate users* will enjoy the benefits of language technology early and at no (or reasonable) cost through a large variety of generic and specialised services offered through a small number of sources. Effective communication requires a unified, high-level, transparent and user-friendly approach with common goals. Especially collaboration between research partners in different countries (both within and outside of Europe) should be eased by providing better visa regulations etc. This also aligns with the request for less complicated management and administration process of EU level.

In order to allow for the broad range of potential solutions, the emerging platforms will have to host (and share) all relevant simple services, including components, tools and data resources, as well as various layers or components of higher services that incorporate simpler ones. Resource exchange infrastructures such as, for example, META-SHARE (including the Linguistic Linked Open Data cloud) will play an important role in the design of the platform. Regarding the setup of the platform and the collaboration between respective stakeholders about 30% of all survey answers emphasise the importance of easy accessibility and open licensing for available tools and data. Commonly agreed upon exchange formats and standards also need to be set up.

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

2.7. Industry and Research

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

To achieve this ambitious plan, all stakeholders need to collaborate and cooperate closely and in a tightly coordinated and efficient way. To demonstrate that the whole Multilingual Europe community firmly stands together, this document is presented by the Cracking the Language Barrier federation, which – at the time of writing (September 2017) – consists of 10

organisations and more than 20 projects already working together on the technological foundations of a Multilingual Europe, some of them from as early as 2010 (META-NET).

Regarding the governance of a potential HLP, one shared suggestion is the wish to put democratic organisation processes in place, e. g. with shifting presidents and elected committee and board members among institutions and countries. Also highlighted was the need to reposition the strategy of EU research with a focus on scientific breakthroughs in order to diversify from the US and large corporation paradigms. This involves fostering strong collaborations between stakeholders, integration of user and customer experience as well as feedback processes, following market-driven approaches to ensure industrial growth.

Specifically targeted Industry policies, as discussed in the STOA study, should foster the creation and growth of competitive European firms while increasing the availability of highly qualified workers. Suggested measures made by the survey respondents to meet these goals: 74% envision closer collaboration between academia and industry (e. g., through job fairs and hackathons). A large percentage of 62% also sees opportunities in the reorganisation of university curriculums, 43% emphasise the importance of fostering a more entrepreneurial culture through specialised course modules, accelerator programmes etc.

2.8. Timeframe and Costs

In terms of duration a majority of the LT community that participated in the survey suggest a time span of 5-15 years (with an inclination towards more than 10 years) as necessary in order to see satisfactory and sustainable results. 7% believe that 5 years is a sufficient period, 35% opt for 5-10 years and another 35% for 10-15 years.

As far as funding is concerned a shared responsibility between the European Union, industry and member states was envisioned with the EU as the stakeholder that should be “naturally” responsible. The distribution of votes for stakeholder involvement looks as follows: European Commission (89%), Industry (57%) and Member states (57%).

The key principle of the Human Language Project is a solid, robust and operational **Multilingual Services** layer, on top of which **Multilingual Applications** can be realised as innovative solutions. We can start with the further design and implementation for the Multilingual Services quite rapidly due to significant investments for the EU into language technologies topics in the last ca. 15 years. Even with a partially established, incomplete set of Multilingual Services, we can start building Multilingual Applications already in 2018. At the same time, we can start bringing in new results from **Research** into the set of Multilingual Services to provide revised, improved or completely new, additional or alternative services. The three-layer-approach does *not* mean that we have to invest in research first to harvest the results later – quite the contrary, i.e., activities on all three layers can be initiated at the same time. In that regard, it is important to note that we anticipate projects not to focus upon one single layer but to address two layers at the same time (**Research and Multilingual Services** or **Multilingual Services and Multilingual Applications**); several larger actions may address all three layers.

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

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

3. Applications and Solutions

In this chapter, we provide a list of multilingual applications and services which is not meant to be exhaustive but constitutes an important set of needed components. It reflects the current state of discussion within the Language Technology Research and Innovation community.

The customers of these solutions are companies (primarily SMEs), research institutes, public administrations, European institutions, the European citizens and other stakeholders interested in the multilingual Digital Single Market.

The answers collected through the survey clearly reveal that a large majority (73%) sees great potential for their current research to make substantial contributions to the improvement of multilingual services for Language Resources and Technologies. Many also see realistic advances for Translation Services (46%) and multilingual solutions for e-Learning (41%), e-Health (38%), and Content Curation (38%) in the near future.

3.1. Area: Multilingual E-Commerce

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

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

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

3.1.2. Multilingual Application: Online Dispute Resolution

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

The EC is devising and deploying an interactive and free-of-charge website for Online Dispute Resolution (ODR). ODR aims at resolving contractual disputes from European consumers (B2C) or traders (B2B), which arise from cross-border and domestic online sales or service

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

3.2. Area: Content, Media, Verticals

3.2.1. Multilingual Application: E-Learning

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

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

3.2.2. Multilingual Application: E-Health

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

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

3.2.3. Multilingual Application: Content Curation and Content Production

- Support the intelligent authoring, enrichment, and processing of content, making it readable and understandable across language linking barriers and for machines and humans alike

- Support the multilingual, automatic or semi-automatic generation of reports and articles based on Big Data, Linked Data and other data sets

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

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

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

The number of connected devices is continuously growing (Internet of Things, Web of Things). Depending on their function and complexity, the nature of desired or needed communication can vary widely. Some objects will come with interesting textual information (manuals, consumer information), others will provide information on their state and will have their own individual digital memory. Objects that can perform actions, such as vehicles and appliances, will accept and carry out (multilingual) voice, gesture or eye-tracking commands. Wearable sensors can provide signals about a person’s mood or emotional state, offering new affect-focused multimodal interaction with devices. In addition, robots are now evolving into collaborative, social machines that will eventually provide useful services to humans in numerous work, medical, educational and household contexts. Dialogue systems and multimodal conversational interfaces that support natural language commands have the potential to adapt automatically to the user and to the environment. For instance, interfaces adapted to elderly people will take into account cognitive, auditory, visual, and articulatory ageing; interfaces will adapt to what a user is doing (working in a noisy, hands-free environment, e.g., when rushing for a train); systems for devices where “traditional” interfaces (keyboard, mouse, trackpad, touchscreen, etc.) are not usable (e.g., small wearable devices) or not appropriate (“companion” systems); smart mobile agents which are capable of deeper natural language and multimodal interaction, possibly focused on specific domains, and capable of rich question answering. Many vertical market sectors with domain-specific assistants exist: shopping, travel, social service planning, learning and tutoring.

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

- Voice of the customer: Multilingual market research, i.e., extracting and interpreting the multilingual opinions with high accuracy, across languages and modalities, and analysing sentiment as well as opinion at deeper levels beyond mere polarity (including intention)

- Voice of the citizen: the same set of technologies but applied to social, sociological, ecological and other related, non-commercial aspects; the goal is to enrich democracy by new mechanisms for improved collective solutions and decisions (also see E-Participation).
- Technically: large-scale, web-scale sentiment analysis, opinion mining, multilingual report generation, trend analysis

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

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

3.2.6. Multilingual Application: E-Government

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

The creation of a multilingual Digital Single Market should also include vastly improved interoperable cross-border public and government services that counter market fragmentation, in particular in the areas of e-government, e-health and e-procurement. This set of solutions foresees, among others, e-procurement platforms in which multilingual language technologies can support the translation of user interfaces, documents and large narratives that are currently performed manually. Language technologies are needed for concept identification and

extraction, matching offer and demand to identify business opportunities and to produce accurate summaries for decision making in tendering processes.

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

3.2.7. Multilingual Application: E-Justice

- Cross-border e-Justice informs citizens, businesses, lawyers and judges about cross-border legal questions and supports mutual understanding of different country-specific legal systems and thereby contributes to the creation of a single European area of justice.

E-justice technology applications give citizens better and improved online access to judicial action. Examples include the settlement of disputes as well as the imposition of criminal sanctions. Every year more than 10 million citizens face judicial procedures involving different EU countries. In 2010 the EU Commission launched the first version of an e-Justice portal as a multilingual online access point that eases life for citizens and businesses in Europe. Citizens can get further information on how the 28 EU countries' legal systems work. The Portal supports them when dealing with events such as divorce, death, litigation, succession etc. Through this portal they can find legal practitioner in another country and learn how to avoid costly court cases through mediation, which country's law applies to their case and whether they are eligible for legal aid.

With more than 10.000 pages of content, the portal provides a great wealth and variety of resources, information and links on laws and practices in all EU countries. This data needs to be structured, organised and interlinked. In addition, language technologies can support the translation and summarisation of documents and provide semantic analyses (e.g. opinion mining etc.) when needed.

New information, tools and functions relying on multilingual technologies are being continuously developed and added. For instance, the Portal will soon feature a Search Engine which will allow legal practitioners to easily find case law as the adoption of the ECLI (European case law identifier) standard is gradually being introduced.

3.3. Area: Translation, Language, Knowledge, Data

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

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

processing and production. A similar approach is currently followed by CEF AT with dedicated service contracts to identify and collect data sets in the CEF-participating countries.

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

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

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

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

3.3.2. Multilingual Application: Translation Centre

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

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

use but it will offer extensive business opportunities for a wide range of service and technology providers.

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

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

High-quality services require a combination of Human Translation (HT), Computer-Assisted Translation (CAT) and full Machine Translation (MT). Core requirements are trust in the reliability and accuracy of translation and the security of the translation channel. The platform should include translation companies and experts in a variety of domains (e.g., bio- medical, financial, legal, scientific), tasks and genres (e.g., technical documentation, business reports, fiction, etc.); extended services like multilingual text authoring, multimedia translation, and quality assurance by experts; mechanisms for customer care and trust building; certified security systems; quality upscale models (instant quality upgrades). In this platform a close collaboration and bridge to the MT@EC system (CEF AT) is foreseen. In addition to the EC MT services, similar services can be provided by private companies, research centres, universities or NGOs such as Translators without Borders or the Rosetta Foundation. The platform could be operated by an industrial interest group (EEIG) in close collaboration with MT@EC. Necessary ingredients are a powerful and stable service and service brokerage platform with an API to automatic or quasi-real time human services provided by a set of initial LSPs and MT systems. It could be hosted by trusted service centres, i.e., certified service providers fulfilling highest standards for privacy, data protection, confidentiality and security of data and translations.

3.3.3. Multilingual Service: Knowledge and Data Repositories

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

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

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

Many elements of the platform are already in place, based on open source tools, specifications and guidelines from the LOD2 stack and the W3C Data Activity. Guidelines and tools specific to linguistic linked data are being actively promoted by several W3C, LT and Knowledge communities. Massively multilingual examples of aggregation and discovery solutions using LLD are publically available and already have a major impact on the NLP and language resource communities. Babelnet[39] aggregates lexical-conceptual information from Wikipedia, Wikidata, different Wordnets and Wiktionary into a single service supported by the annotation service Babelfy. It covers 271 languages, 117 million lexical senses, over 6 million concepts, 7 million named entities, 10 million images, all interlinked by 354 million lexico-semantic relations using nearly 2 billion RDF triples. The Linguistic Linked Data Cloud[40] of interlinked resources is now an important and growing part of the overall LOD cloud. The language resource community is well on the way to wholesale adoption of linked data as its primary data exchange mechanism.

[39] http://babelnet.org/
[40] http://linguistic-lod.org/llod-cloud

4. Research Themes

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

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

4.1. Research Theme: Cross-lingual Big Data Language Analytics

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

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

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

4.1.1. Novel Research Approaches and Targeted Breakthroughs

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

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

- Semantification of the web: bridging between the semantic islands and the traditional web containing unstructured data.
- Integration of textual and multimedia data with social network and social media data on dimensions including semantics, context, location and temporal; this integration needs to be made possible with regard to typical data representations, i.e., big, heterogeneous, distributed, user-tagged, user-generated, multimodal; the representation has to be lightweight and augmented with semantics including the extraction of semantic representations and transforming them into representations for reasoning and inferencing.
- Aligning and making different types and genres of content (e.g., news, social media, blogs, academic texts, archives etc.) comparable as well as interoperable.

- The methods need to be able to operate not only on the actual linguistic data (both spoken language and written texts) but also take into account available metadata, arbitrary markup and other annotations as well as multimedia data (including video, images, audio).
- Among the specific targeted breakthroughs are the following: detecting and monitoring opinions, demands, needs and economic as well as social issues; detecting diversity of views, biases along different dimensions (e.g., demographic) including temporal (evolution of opinions); support for decision makers and communication participants; problem mining and problem solving; support of collective deliberation and collective knowledge accumulation; vastly improved approaches to sentiment detection and scoring; genre-aware text and language processing; topic recommendation; understanding content and influence diffusion in open and closed social media (identifying drivers of opinion spreading).
- The processing of vast amounts of content also makes it necessary to increase research in the retrieval and summarisation of heterogeneous content sources, e.g., passage retrieval to support question-answering tasks, query rewriting methods and multi-document summarisation with both shallow and deep techniques (including multimodal).
- As a complement, novel approaches are needed with regard to text generation, especially the generation of written and spoken language content based on extracted knowledge (semantic storytelling) or based on existing data sets (data-to-text).
- Of specific importance are sophisticated methods for topic and event detection that are tightly integrated with the Semantic Web and Linked Open Data, especially multimodal clustering approaches based on heterogeneous features; automatic and user-driven clustering; clustering and classification based on semantic representations; event detection in multimedia content by exploiting semantic and textual features from speech recognition and captions, as well as visual and motion information.
- In terms of decision support techniques, research is to be intensified on semantic reasoning (rule based, fuzzy, backward and forward chaining) that operates on semantically integrated data, supervised models trained with a variance of features (e.g., concepts, name entities, n-grams, contextual characteristics, sentiment), visual analytics.
- With regard to the validation and gathering of provenance information, new methods are needed for the detection of fake content, identification of contradictory facts and hidden relations including repeated or similar facts along the spatiotemporal axis.

4.1.2. Solution and Realisation

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

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

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

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

4.2. Research Theme: High-Quality Machine Translation

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

4.2.1. Novel Research Approaches and Targeted Breakthroughs

Although recent progress has already led to new applications of MT technology, radically novel and radically different approaches are needed to accomplish the ambitious goal of this research, i.e., a genuine quality breakthrough. Among these new research approaches are a stronger focus on producing high-quality, publishable outbound translations, needed for the success of MT in the language industry.41 Research needs to systematically concentrate on the barriers that still prohibit high-quality translations. For this, a fully implemented, unified, dynamic, weighted, and multidimensional quality assessment model with task and language profiling needs to be devised and adopted by the whole research community. This also includes improved automatic quality estimations for given task specifications and the inclusion of translation professionals in the research and innovation process. Vice versa, human translation needs to be enhanced with ergonomic, computer-supported work environments and multilingual text authoring. The recent breakthroughs in neural MT need to be improved through additional statistical models that extract more dependencies from the data. In addition, a semantic translation paradigm is needed by extending statistical translation with semantic data such as linked open data, ontologies including semantic models of processes and textual inference models. We also want to put a stronger emphasis on the properties of individual

[41] 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.

languages, especially through the exploitation of strong monolingual analysis as well as generation methods and resources. Furthermore, we want to intensify research on modular combinations of specialised analysis, generation and transfer models, permitting accommodation of registers and styles (including user-generated content) and also enabling translation within a language (e.g., between specialists and laypersons).

The expected breakthroughs will include high-quality text translation and reliable real-time speech translation for all official European languages as well as regional and minority languages; a modular analysis-transfer-generation translation technology that facilitates reuse and constant improvement of (statistical and knowledge-driven) modules; automatic subtitling and voiceover of films and multimedia applications in selected domains, such as public service, sports events, and other applications; always-correct translation for critical subdomains.

4.2.2. Solution and Realisation

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

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

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

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

to be built include trusted service clouds, workbenches for translators and translation workflows.

4.3. Research Theme: Meaning, Semantics, Knowledge

While Machine Translation is a key technology for the Multilingual Digital Single Market, other technologies are needed to tailor engagement between companies and their customers to the domain being addressed. Customers should be able to search for user-generated content (UGC) on a specific product or service regardless of the language in which it was posted. Image, video and audio postings on products should be tagged, summarised, discoverable and accessible to users in any other language. Customer profiles should be built in their native language so the personalisation of engagement can be automated in that language, while still providing market intelligence in the vendor's native language. To successfully tailor such cross- and multilingual customer experiences, companies must monitor and analyse UGC on social media, blogs, forums and product review sites and react continuously with well targeted customer engagement. The effectiveness of Language Technologies is, however, limited by the distance between the linguistic data available to train them and the content they must process when deployed in a specific application. This is especially problematic for SMEs. Small companies succeed by excelling in a specific niche where they must engage skilfully with their customers using and understanding the terms and language patterns specific to that niche. One-size-fits-all language technologies, such as unrestricted machine translation, will fail to meet the language needs of specialised SMEs. Small companies, however, typically lack the knowledge or capacity to assemble their own linguistic data assets to tailor language technology to their needs. Without tailored language technology support, though, SMEs will not be able to make use of the DSM because of the language barriers to bidirectional customer engagement. Linguistic Linked Data is already proving a scalable source of massively multilingual open language resources for LT services. Research is needed into tools and techniques to integrate the lifecycle management of linguistic data into technologies that apply to the specific niches of online discourse that SMEs must use. SMEs must be empowered with cheap and easy tools to assemble, deploy and refine micro-domains for linguistic and semantic resources that can be used across different LT components they employ.

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

4.3.1. Novel Research Approaches and Targeted Breakthroughs

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

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

Generating insights from unstructured data: Data is often in unstructured form, such that it cannot be directly exploited in applications or to generate insights. Robust, efficient and scalable techniques for refining unstructured data in such a way that it can be transformed to

make it usable are needed. Human language technologies and NLP methods play a crucial role and need to be extended in terms of coverage, robustness and scalability.

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

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

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

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

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

4.3.2. Solution and Realisation

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

In the Jeopardy game show, IBM's Watson was able to find correct answers that none of its human competitors could provide, which might lead one, erroneously, to think that the problem of automatic question answering is solved. With clever lookup and selection mechanisms for the extraction of answers, Watson could actually find the right responses without a full analysis of the questions from a huge set of handbooks, decades of news, lexicons, dictionaries, bibles, databases, and the entire Wikipedia. Outside the realm of quiz shows, however, most questions that people might ask cannot be answered by today's technology, even if it has

access to the entire web, because they require a level of language and context understanding that is not yet possible with today's technology. Modelling the contexts in which users ask questions must therefore be efficiently indexed against into the increasingly massive body on multilingual knowledge from which answers can be sourced.

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

4.4. Research Theme: Conversational Technologies

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

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

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

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

4.4.1. Novel Research Approaches and Targeted Breakthroughs

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

meaning, style and emotion. They also have to be equipped with methods for incremental speech, including pauses and hesitations.

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

A common push has to be made towards more natural dialogue. This includes, among others, the recognition and production of paralinguistics (prosody, visual cues, emotion) and a better understanding of socio-emotional functions of communicative behaviour, including group dynamics, reputation and relationship. In addition, more natural dialogue needs more advanced dialogue models that are proactive (not only reactive), that are able to detect that recognised speech is intended as a machine command, they have to be able to interpret silence as well as direct and indirect speech acts (including lies and humour). Another prerequisite for more natural dialogue is the ability of the system to personalise itself to the user's preferences. The system has to operate in a transparent way and be able to participate in multi-party conversations and make use of other sensory data (GPS, RFID, cameras etc.).

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

4.4.2. Solution and Realisation

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

5. Horizontal Topics

In this chapter, we briefly discuss several horizontal aspects of the Human Language Project.

5.1. Standardisation and Interoperability

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

5.2. Business Models and Ecosystems

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

- **2020:** Enable localisation industry to explore new business models, beyond translation, and contributing, e.g., to marketing of digital content
- **2020:** Build a connection between creators, distributors and consumers of public sector information (PSI), to allow for feedback on the usefulness of public data sets in new business models

5.3. Language Policies and Public Procurement

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

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

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

The search for originality on the side of the public research funding bodies and their constant trial-and-error search for new topics that might finally help the European IT industry to be in time with their innovations have often caused the premature abortion of promising developments, whose preliminary results were more than once taken up by research centres and enterprises in the US. An example in language technology is the progress in statistical

machine translation. Much of the groundwork laid in the German government-sponsored project Verbmobil (1993–2000) was later taken up by DARPA research and commercial systems – including Google Translate.

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

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

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

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

5.4. Copyright and Data Protection

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

5.5. Open Source

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

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

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

5.6. Related Areas, Applications and Societal Challenges

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

There is a close relationship to interactive and multilingual spoken language interfaces and robots (especially the SPARC Robotics PPP), connected machines (Advanced Manufacturing, Industry 4.0), Inclusion, E-Learning as well as generic connected devices (Internet of Things, Web of Things). The relationship between multilingual technologies and ecommerce applications is so evident and of such vital importance that we also mention this area, as well as the emerging trend to Smart Cities and Smart Services.

The importance of the languages in our European society has never been in the focus of attention as compared to other highly multilingual societies like South Africa or India where language borders hinder exchange and communication *within* a state. According to the principles of the UN-endorsed World Summit on the Information Society, the "Information Society should be founded on and stimulate respect for cultural identity, cultural and linguistic diversity." Recent scientific work has shown that even our moral decisions are influenced by whether we are speaking our mother tongue or a foreign language.[42]

In fact, the technology solutions detailed in the next chapter address many of the societal challenges specifically to be taken into account by activities under the framework of Horizon 2020.[43] The following list provides several examples:

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

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

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

6. Conclusions

6.1. Expected Economic Impact

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

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

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

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

6.2. Potential Funding Sources

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

The European Union could initially support the HLP especially through dedicated activities in upcoming Horizon 2020 calls (2018–2020) and through Connecting Europe Facility (CEF). Horizon 2020 Research Actions are compatible to our planned activities in the area of research, while Horizon 2020 Research and Innovation Actions as well as Coordination and Support Actions are needed for the actual innovation and deployment activities. 25 million Euros will be available for Language Technology projects in call ICT-29-2018, "A multilingual Next Generation Internet". Highly innovative activities with a major commercial impact are needed for the Application Areas – especially here, the European language technology industry will participate (most of these companies are SMEs). Through CEF, deployment and

innovation actions could be funded, especially with regard to public online services. Furthermore, there are horizontal programmes such as Horizon 2020 Widespread/Teaming that could boost the knowledge and technology transfer between countries that already have excellent research and innovation hubs in language technology and those that do not; the goal would be to enable the less innovative countries to develop technologies for their respective languages. Similar programmes to boost SMEs exist.

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

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

Appendix

A. Editorial Team

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

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

Representatives from the Cracking the Language Barrier federation:
Núria Bel, Kalina Bontcheva, John Judge, Maite Melero, Steve Renals,
Felix Sasaki, Andrejs Vasiljevs

B. History of this Document

Version 0.5: Second version, after a preliminary draft version; Version 0.5 was presented at META-FORUM 2015 and Riga Summit 2015 (April 2015).

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

Version 1.0 beta: Presented and discussed at META-FORUM 2017 (November 2017).

Version 1.0: Published in December 2017.

C. Input Documents

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

- Abney, S., & Bird, S. (2010). The human language project: Building a universal corpus of the world's languages. In Proceedings of the 48th annual meeting of the association for computational linguistics (pp. 88–97). Uppsala: Association for Computational Linguistics.
- Philipp Cimiano (2015): "The LIDER Roadmap in a nutshell", presentation given at the Workshop on multilingual data value chains in the Digital Single Market, 16 January 2015, Brussels, Belgium. https://ec.europa.eu/digital-agenda/en/news/workshop-multilingual-data-value-chains-digital-single-market.
- Gerald Cultot (2015): "eHealth services – multilingual challenges", presentation given at the Workshop on multilingual data value chains in the Digital Single Market, 16 January 2015, Brussels, Belgium. https://ec.europa.eu/digital-agenda/en/news/workshop-multilingual-data-value-chains-digital-single-market.
- Andrew Joscelyne (late 2014): "A Strategic Research and Innovation Agenda for a Conversational European Digital Marketplace" (draft position paper).
- Nils Lenke (2015): "Nuance Inc.", DFKI Tech Day, 30 January 2015, DFKI Saarbrücken, Germany.
- Dave Lewis (2015): "Shopping Across the Language Barrier", presentation given at the Workshop on multilingual data value chains in the Digital Single Market, 16 January 2015, Brussels, Belgium. https://ec.europa.eu/digital-agenda/en/news/workshop-multilingual-data-value-chains-digital-single-market.
- LIDER (10/2014): "Roadmap for the use of Linguistic Linked Data for content analytics"
- META-NET (2013): "Strategic Research Agenda for Multilingual Europe 2020", Georg Rehm and Hans Uszkoreit (eds.), presented by the META Technology Council. Springer.
- MLi (09/2014): "D5.1 – Big and Social Language Data Requirements for the MLi Hub".

- QTLaunchPad (11/2014): "European Quality Translation Research 2015: Ongoing Work and Roadmap".
- Ruben Riestra (2015): "Multilingual data value chains in the Digital Single Market", report presented at the Workshop on multilingual data value chains in the Digital Single Market, 16 January 2015, Brussels, Belgium. https://ec.europa.eu/digital-agenda/en/news/workshop-multilingual-data-value-chains-digital-single-market.
- ROCKIT (10/2014): Roadmap for Conversational Interaction Technologies – Coordination and Support Action, D1.1 Innovation Drivers, future scenarios, and best practice.
- ROCKIT (10/2014): Roadmap for Conversational Interaction Technologies – Coordination and Support Action, D2.1 First Report on Innovation in the ROCKIT Domain.
- ROCKIT (10/2014): Roadmap for Conversational Interaction Technologies – Coordination and Support Action, D3.1 First Report on Research in the ROCKIT Domain.
- ROCKIT (02/2014): Roadmap for Conversational Interaction Technologies – Coordination and Support Action, D4.1 ROCKIT Roadmap Specifications.
- Alan Mas Soro: "Language Technologies for Europe: A way to foster SME internationalization", presentation given at the Workshop on multilingual data value chains in the Digital Single Market, 16 January 2015, Brussels, Belgium. https://ec.europa.eu/digital-agenda/en/news/workshop-multilingual-data-value-chains-digital-single-market.
- STOA. (2017). Language equality in the digital age – Towards a Human Language Project. STOA study (PE 598.621), IP/G/STOA/FWC/2013-001/Lot4/C2, March 2017. Carried out by Iclaves SL (Spain) at the request of the Science and Technology Options Assessment (STOA) Panel, managed by the Scientific Foresight Unit (STOA), within the Directorate-General for Parliamentary Research Services (DG EPRS) of the European Parliament, March. http://www.europarl.europa.eu/stoa/.
- Adomas Svirskas (2015): "Pan-European Electronic Document Platform. Open Interoperable Solution for Europe", presentation given at the Workshop on multilingual data value chains in the Digital Single Market, 16 January 2015, Brussels, Belgium. https://ec.europa.eu/digital-agenda/en/news/workshop-multilingual-data-value-chains-digital-single-market.
- Hans Uszkoreit (2014): "European Platform(s) for Machine Translation and other Language Technologies", presentation given at the META-NET Platform Strategy Meeting during the Language Resources and Evaluation Conference (LREC), 26-31 May 2014, Reykjavik, Iceland.
- Xenios Xenophontos (2015): "Online Dispute Resolution Platform – Multilingual challenges", presentation given at the Workshop on multilingual data value chains in the Digital Single Market, 16 January 2015, Brussels, Belgium. https://ec.europa.eu/digital-agenda/en/news/workshop-multilingual-data-value-chains-digital-single-market.
- Sonja Zillner (2015): "cPPP Big Data Value-SRIA", presentation given at the Workshop on multilingual data value chains in the Digital Single Market, 16 January 2015, Brussels, Belgium. https://ec.europa.eu/digital-agenda/en/news/workshop-multilingual-data-value-chains-digital-single-market.

D. Digital Language Extinction in Europe

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

Europe must take action to prepare its languages for the digital age. They are a precious component of our cultural heritage and, as such, they deserve future-proofing. The META-NET

study shows that, in the digital age, multilingual Europe and its linguistic heritage are facing challenges but also many possibilities and opportunities.

The study, prepared by more than 200 experts and documented in 31 volumes of the META-NET White Paper Series (available both online and in print), assessed language technology support for each language in four different areas: automatic translation, speech interaction, text analysis and the availability of language resources. A total of 21 of the 30 languages (70%) were placed in the lowest category, “support is weak or non-existent” for at least one area by the experts. Several languages, for example, Icelandic, Lithuanian and Maltese, receive this lowest score in all four areas but it must be noted that support for some of the languages with smaller numbers of speakers is slowly increasing since the original publication of the META-NET White Paper Series in 2012. At the other end of the spectrum, while no language was considered to have “excellent support”, only English was assessed as having “good support”, followed by languages such as Dutch, French, German, Italian and Spanish with “moderate support”. Languages such as Basque, Bulgarian, Catalan, Greek, Hungarian and Polish exhibit “fragmentary support”, placing them also in the set of high-risk languages.

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

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

* (including online and public services)

Unified Customer Experience

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

Voice of the Customer and Voice of the Citizen

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

Digital Translation Centre

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

Content Curation and Production

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

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

This Strategic Agenda (including previous versions) has received funding from the EU's Horizon 2020 research and innovation programme under grant agreements No. 645357 (CRACKER) and No. 644583 (LT_Observatory).

4 LREC 2018 Paper: Language Technology for Multilingual Europe: An Analysis of a Large-Scale Survey regarding Challenges, Demands, Gaps and Needs

In the following, we include an extended abstract titled "Language Technology for Multilingual Europe: An Analysis of a Large-Scale Survey regarding Challenges, Demands, Gaps and Needs", written by Georg Rehm and Stefanie Hegele (DFKI). This extended abstract was submitted to the conference LREC 2018, which is to take place in Miyazaki, Japan, in May 2018. The abstract has already been positively reviewed and was accepted for publication and presentation at the conference.

Language Technology for Multilingual Europe: An Analysis of a Large-Scale Survey regarding Challenges, Demands, Gaps and Needs

Georg Rehm, Stefanie Hegele
DFKI GmbH, Alt-Moabit 91C, 10559 Berlin, Germany
Corresponding author: georg.rehm@dfki.de

Abstract

We present the analysis of a large-scale survey titled "Language Technology for Multilingual Europe" and conducted between May and June 2017. A total of 634 participants in 52 countries participated in the survey. Its main purpose was to collect input, feedback and ideas from the European Language Technology research and innovation community in order to assess the most prominent research areas, projects and applications, but, more importantly to identify the biggest challenges, obstacles and gaps Europe is currently facing with regard to its multilingual setup and technological solutions. Participants were encouraged to share concrete suggestions and recommendations on how present challenges can be turned into opportunities in the context of a long-term, large-scale, Europe-wide research, development and innovation funding programme, currently titled Human Language Project.

Keywords: Multilinguality, LR National/International Projects, Infrastructural/Policy issues, LR Infrastructures and Architectures

1. Introduction

Europe is a multilingual society with 24 official member state languages and many additional unofficial and regional languages as well as languages of minorities, immigrants and important trade partners. Nevertheless, day in and day out, language barriers keep severely hampering the free flow of information, thought, ideas, goods and products through the continent. Powerful multilingual as well as cross-lingual and monolingual language technologies, making use of the latest Artificial Intelligence algorithms in combination with ever-growing data sets, have the potential of helping to overcome language barriers.

The recent study "Language Equality in the Digital Age – Towards a Human Language Project", commissioned by the European Parliament's Science and Technology Options Assessment Committee (STOA), recommends, to the European Union, to initiate a new, large-scale European Language Technology research, development and innovation flagship programme, called, in the study, the *Human Language Project* (HLP) (STOA, 2017). It is foreseen to be a long-term European collaborative programme between research, innovation, industry, academia, administrations and citizens with the goal of achieving the next scientific breakthroughs for the automatic processing and generation of written or spoken natural language. In addition to basic research, the HLP is foreseen to include applied research as well as innovation and commercialisation activities. More concrete details on the uniqueness of the HLP, which needs to be specifically designed for Europe's demands, are discussed in section 4.. A key goal of our survey was to get an overview of the current situation of Language Technology research activities throughout Europe and to determine where important gaps and obstacles exist.

2. Recent Developments

The principle that all 24 official languages share an equal status and are supported on the same level is perpetuated in the EU Charter (Article 22) as well as in the Treaty on the European Union (Art. 3(3) TEU). The META-NET White Paper Series, however, has revealed that there is a steadily increasing and rather severe threat of digital extinction for at least 21 European languages (Rehm and Uszkoreit, 2012; Rehm et al., 2014).

To address this threat and recognise Europe's opportunities, among others, in the fostering of a truly Digital Single Market, META-NET[1] (a Network of Excellence consisting of more than 60 research centers in 34 European countries) has been committed to support work on multilingual technologies and to provide strategic guidance since 2010 (Rehm and Uszkoreit, 2013; Rehm et al., 2016b; Rehm et al., 2016a). META-NET is currently being supported and funded through the EU project CRACKER.[2] CRACKER's objectives encompass, among others, the publishing of research and innovation agendas (Rehm, 2015; Rehm, 2016; Rehm, 2017). It has also established the Cracking the Language Barrier[3] federation which acts as an umbrella initiative for European projects and organisations working on technologies for multilingual Europe.

Europe has a long-standing research, development and innovation tradition with several hundred universities and research centers performing excellent, highly visible and internationally recognised research on all European and many non-European languages. Especially in the field of Machine Translation most of the basic research has happened in European research projects. Moses (Koehn et al., 2007), until 2016 the state of the art phrase-based statistical MT system, and recent European NMT results, especially those of the European research project QT21, are just two examples for excellence and world class research (Bojar et al., 2017). Nonetheless, challenges are omnipresent and must be addressed by the EU, the Member States as well as stakeholders from academia and industry.

3. Method

The survey contains a total of 29 questions (see Appendix A) of which 16 are open questions with free text answers. The remaining ones are a mixture of multiple choice and yes/no

[1] http://www.meta-net.eu
[2] http://www.cracker-project.eu
[3] http://www.cracking-the-language-barrier.eu

– extended abstract submitted to LREC 2018 – accepted for publication – please don't circulate –

Figure 2: First page of the LREC 2018 paper "Language Technology for Multilingual Europe: An Analysis of a Large-Scale Survey regarding Challenges, Demands, Gaps and Needs"

Language Technology for Multilingual Europe: An Analysis of a Large-Scale Survey regarding Challenges, Demands, Gaps and Needs

Georg Rehm, Stefanie Hegele
DFKI GmbH, Alt-Moabit 91C, 10559 Berlin, Germany
Corresponding author: georg.rehm@dfki.de

Abstract

We present the analysis of a large-scale survey titled "Language Technology for Multilingual Europe" and conducted between May and June 2017. A total of 634 participants in 52 countries participated in the survey. Its main purpose was to collect input, feedback and ideas from the European Language Technology research and innovation community in order to assess the most prominent research areas, projects and applications, but, more importantly to identify the biggest challenges, obstacles and gaps Europe is currently facing with regard to its multilingual setup and technological solutions. Participants were encouraged to share concrete suggestions and recommendations on how present challenges can be turned into opportunities in the context of a long-term, large-scale, Europe-wide research, development and innovation funding programme, currently titled Human Language Project.

Keywords: Multilinguality, LR National/International Projects, Infrastructural/Policy issues, LR Infrastructures and Architectures

1. Introduction

Europe is a multilingual society with 24 official member state languages and many additional unofficial and regional languages as well as languages of minorities, immigrants and important trade partners. Nevertheless, day in and day out, language barriers keep severely hampering the free flow of information, thought, ideas, goods and products through the continent. Powerful multilingual as well as cross-lingual and monolingual language technologies, making use of the latest Artificial Intelligence algorithms in combination with ever-growing data sets, have the potential of helping to overcome language barriers.

The recent study "Language Equality in the Digital Age – Towards a Human Language Project", commissioned by the European Parliament's Science and Technology Options Assessment Committee (STOA), recommends, to the European Union, to initiate a new, large-scale European Language Technology research, development and innovation flagship programme, called, in the study, the *Human Language Project* (HLP) (STOA, 2017). It is foreseen to be a long-term European collaborative programme between research, innovation, industry, academia, administrations and citizens with the goal of achieving the next scientific breakthroughs for the automatic processing and generation of written or spoken natural language. In addition to basic research, the HLP is foreseen to include applied research as well as innovation and commercialisation activities. More concrete details on the uniqueness of the HLP, which needs to be specifically designed for Europe's demands, are discussed in section 4.. A key goal of our survey was to get an overview of the current situation of Language Technology research activities throughout Europe and to determine where important gaps and obstacles exist.

2. Recent Developments

The principle that all 24 official languages share an equal status and are supported on the same level is perpetuated in the EU Charter (Article 22) as well as in the Treaty on the European Union (Art. 3(3) TEU). The META-NET White Paper Series, however, has revealed that there is a steadily increasing and rather severe threat of digital extinction for at least 21 European languages (Rehm and Uszkoreit, 2012; Rehm et al., 2014).

To address this threat and recognise Europe's opportunities, among others, in the fostering of a truly Digital Single Market, META-NET[1] (a Network of Excellence consisting of more than 60 research centers in 34 European countries) has been committed to support work on multilingual technologies and to provide strategic guidance since 2010 (Rehm and Uszkoreit, 2013; Rehm et al., 2016b; Rehm et al., 2016a). META-NET is currently being supported and funded through the EU project CRACKER.[2] CRACKER's objectives encompass, among others, the publishing of research and innovation agendas (Rehm, 2015; Rehm, 2016; Rehm, 2017). It has also established the Cracking the Language Barrier[3] federation which acts as an umbrella initiative for European projects and organisations working on technologies for multilingual Europe.

Europe has a long-standing research, development and innovation tradition with several hundred universities and research centers performing excellent, highly visible and internationally recognised research on all European and many non-European languages. Especially in the field of Machine Translation most of the basic research has happened in European research projects. Moses (Koehn et al., 2007), until 2016 the state of the art phrase-based statistical MT system, and recent European NMT results, especially those of the European research project QT21, are just two examples for excellence and world class research (Bojar et al., 2017). Nonetheless, challenges are omnipresent and must be addressed by the EU, the Member States as well as stakeholders from academia and industry.

3. Method

The survey contains a total of 29 questions (see Appendix A) of which 16 are open questions with free text answers. The remaining ones are a mixture of multiple choice and yes/no

[1] http://www.meta-net.eu
[2] http://www.cracker-project.eu
[3] http://www.cracking-the-language-barrier.eu

questions. The survey is divided into three main parts covering (1) background, research interests and projects of the participants, (2) visions for a large-scale European Language Technology research and development programme and (3) ideas on talent generation and retention in Europe. Participants were not obliged to answer all questions, but encouraged to fill in the ones they feel comfortable with. The survey was designed and set up using the service Typeform[4], a software for building online forms.

The survey was launched on 16 May and closed on 4 July 2017. As an incentive to maximise the number of answers, those who submitted the survey had the chance to win a tablet computer. After testing and making sure that the questions were phrased the right way, the survey was shared within a smaller circle (mainly members of META-NET, META, CRACKER as well as members of the Cracking the Language Barrier federation) with an appeal to share the survey within their own respective networks and also through social media. In a second round a wider audience of more than 4000 people was targeted, including participants of former META-FORUM and other conferences as well as respondents of the META-NET Open Letter campaign, conducted in 2015 (Rehm et al., 2016a).[5] We also announced the survey on the major mailing lists.

The survey created a total of 634 responses and, considering the number of questions, a surprisingly high completion rate of 27%. The average time needed for completing the survey was 35,48 minutes (see Figure 1). Both the completion rate and the average time indicate that the respondents are very passionate about the language topic and Europe's multilingual challenge. One major goal of this survey was to bring the European LT community together and gather responses from a wide and demographically distributed audience.

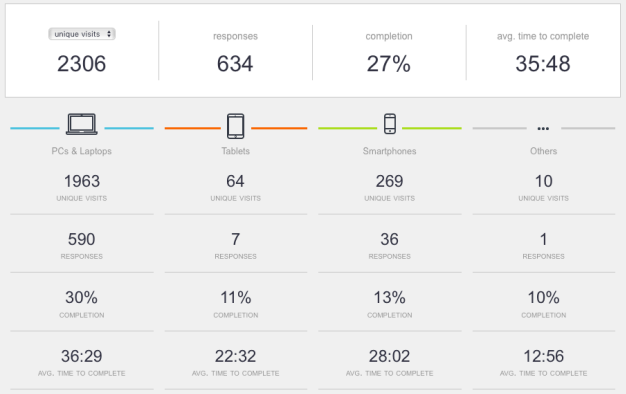


Figure 1: Survey completion rates on different devices

4. Analysis

Since the scope of this extended abstract does not allow to analyse all 29 questions in detail, we focus on the ones we consider most insightful and provide relevant quantitative and qualitative statistics and findings (see Figure 2 for some of the key insights). We refer to specific questions, listed in Appendix A, using abbreviations in the form of Q1, Q2 etc. The analysis follows the survey's original tripartition.

[4]https://www.typeform.com
[5]http://multilingualeurope.eu

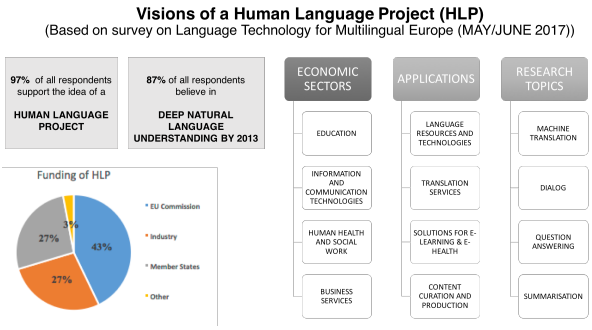


Figure 2: Overview of survey's key input on a HLP

4.1. Part 1: Background and Research Interests

Below we present the demographic details, the current challenges and gaps in terms of technology as well as their economic impact, especially with regard to the Digital Single Market (DSM).

4.1.1. Respondents Demographics

Access statistics of the survey web page and Google Analytics[6] reveal that the survey was opened by potential respondents in 67 different countries. Completed surveys were collected from 52 countries (see Figure 3).

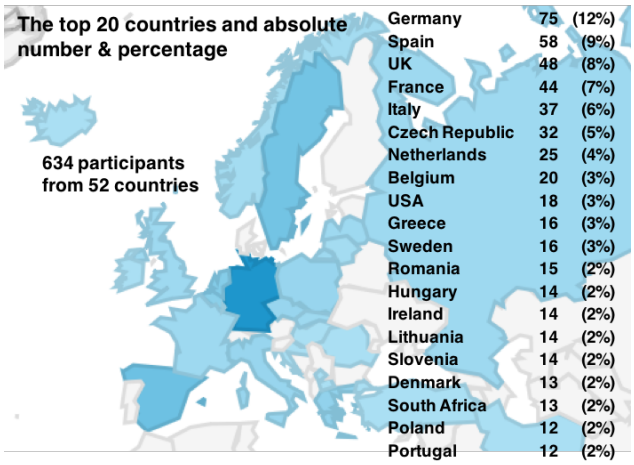


Figure 3: Number of collected responses sorted by country

As for socioeconomic statistics, the distribution shows that a large majority of participants hold senior roles at their respective organisations (such as professor, senior researcher, group leader etc.). This information about the roles seen in context with the seniority level (53% have more than 20 years of work experience and another 27% more than 10 years) and the participation from 52 countries clearly portrays a wide and diverse range of the European Language Technology research and innovation community (and even beyond). This expertise and long experience are accordingly reflected in the high quality of answers collected. The most commonly represented research fields include Language Technology, Computational Linguistics, General Linguistics, Artificial Intelligence and Computer Science.

[6]https://analytics.google.com

While the majority of participants is based at universities and research centers, the survey also reached a substantial group of participants from industry, 33 (corresponding to 5% of all respondents) from large enterprises such as Microsoft, IBM, Intel and Nuance and 68 (11%) from SMEs.

4.1.2. Technological Gaps and Challenges

Regarding crucial gaps in terms of technologies for specific languages (Q13), almost 40% of all respondents highlight that there is insufficient research being done for minority languages and dialects, directly resulting in a shortage of available resources. This lack becomes most evident in Machine Translation applications for smaller European languages as well as other standard NLP tools and systems (according to approx. 19%). Further gaps mentioned are imposed by limited funding for low-resourced languages and copyright restrictions for certain data sets. Further, interoperability and standardisation need to be intensified.

When asked about the biggest challenge the European Language Technology field is facing at the moment (Q14) around 16% of all provided survey answers stress that the neglect of smaller languages is a severe threat, which is leading to a fragmented rather than a united and multilingual Europe. Around 90% state that they work with English in their research (not exclusively though) since they are often given little incentive to solely focus on smaller or minority languages. For instance, when it comes to publishing research results there is a strong bias towards incorporating results for English. Other challenges include the insufficient amount of data resources (approx. 13%), an unwillingness of collaboration within the community (approx. 8%) and, as already indicated above, a lack of funding (approx. 8%).

4.1.3. Economic Impact and the DSM

We also asked the respondents questions (Q11, Q12) regarding the economic impact of language technologies, especially in the context of the Digital Single Market (DSM).

Identified as the sectors to have the highest potential contribution to commercial growth are Education (71%), Information & Communication Technologies (64%) as well as Human Health & Social Work (45%). Specific services and applications that could benefit the Multilingual Digital Single Market comprise better Language Resources and Technologies (73%), Translation Services (46%), Multilingual Solutions for E-Learning (41%) and E-Health (38%). In the context of industries, sectors and verticals the necessity of an on-going knowledge transfer and effective collaboration between academia and industry is highlighted. The Health sector is unequivocally the most significant one, Education comes in second, closely followed by Tourism and Travel as well as Law and Justice.

4.2. Part 2: Visions for a Future Large-Scale Language Technology Programme

In this part we analyse the questions on the organisational set up and governance of a potential Human Language Project, the most important research areas as well as applications and services that should be components of a HLP.

4.2.1. Organisational Frame and Governance

The overall suggestion to initiate a large-scale Human Language Project (HLP) received substantial support from the group of respondents with 97% stating that they are in favour of establishing such a funding programme. Only a very small number of participants (3%) does not agree; their main arguments are unsuccessful previous attempts of similar programmes which did not achieve their targeted goals because of bureaucratic hurdles and a lack of focus. Furthermore, 97% consider the survey's suggested key strategic vision – to achieve Deep Natural Language Understanding and Generation by 2030 – as realistic and therefore an adequate scientific challenge. An appropriate timeframe would be likely to fall in the range of 10-15 years (7% believe that 5 years is a sufficient period, 35% opt for 5-10 years and another 35% for 10-15 years).

As far as funding is concerned a shared responsibility between the European Union, industry and member states was envisioned with the EU as the stakeholder that should be "naturally" responsible. The distribution of votes for stakeholder involvement looks as follows: European Commission (89%), Industry (57%) and Member states (57%).

When it comes to strategic guidance what can be derived from the survey responses is the strong suggestion to concentrate funding on smaller scale projects, starting bottom-up with smaller goals, and also to avoid heavy bureaucracy.

Regarding the governance of a potential HLP, one shared suggestion is the wish to put democratic organisation processes in place, e. g. with shifting presidents and elected committee and board members among institutions and countries. Also highlighted was the need to reposition the strategy of EU research with a focus on scientific breakthroughs in order to diversify from the US and large corporation paradigms. This involves fostering strong collaborations between stakeholders, better school and especially university education with more incentives for young researchers (see Section 4.3.), integration of user and customer experience as well as feedback processes, following market-driven approaches to ensure industrial growth.

4.2.2. Key Research Areas

In terms of research, the Human Language Project aims to tightly intertwine basic research, applied research, innovation and commercialisation (Q20).

As far as basic research is concerned a majority mentioned the further development of existing resources (incl. corpora, ontologies, dictionaries etc.) and improvement of data annotations (approx. 9%). In this context, effective legal frameworks for better accessibility are also necessary. Besides, basic research should be centered around deep learning and neural networks (approx. 7%) as well as Natural Language Understanding (approx. 7%). A majority also highlighted the need to further work on existing NLP tasks and tools such as Question Answering, Summarisation, Information Extraction and Sentiment Analysis (approx. 6%).

Applied research should strongly focus on MT according to around 13% of all respondents. Seen as crucial is thereby, again, the improvement of multilingual resources, data sets and terminology repositories, allowing for standardisation and interoperability (approx. 10%). In addition, there is a

demand for improved open-source platforms with a wide range of available systems and applications and truly open and unencumbered data and code repositories (approx. 4%), which are further discussed in Section 4.2.3..
When it comes to innovation the inclusion of all languages and fostering of inter-cultural systems is regarded as a top priority (9%). This also presupposes better and stronger relations between academia and industry (7%). Also stressed is the need to bring together knowledge and methods developed for different fields and domains, e. g., e-health, e-government and e-justice (5%). In addition, there is an interest for more advanced visualisations and interfaces, new innovative tools incorporating NLU and seamless human-computer as well as human-robot interactions (5%).

4.2.3. Applications and Platforms

As for the most important topics, applications and platforms to be integrated (Q24), Machine Translation is uncontroversially the most important one according to approx. 14% of respondents. Considered as almost equally important are the availability of download services for multilingual resources including ontologies, lexicons, dictionaries etc. (approx. 10%). As for further applications a more in-depth development of already existing NLP tools is encouraged, especially speech applications (approx. 10%). Other listed applications include information extraction and retrieval, summarisation, search systems and intelligent assistants.
Among the topics and domains most relevant for the development of future applications and services are education, health, e-participation and e-government (10%).
Regarding the setup of a European Language Technology Data and Service platform and the collaboration between respective stakeholders (Q25), about 30% of all survey answers emphasise the importance of easy accessibility and open licensing for available tools and data. Commonly agreed upon exchange formats and standards also need to be set up. Almost 11% see an involvement of all stakeholders, i. e., data providers, LT providers and LT consumers, as necessary. Effective communication requires a unified, high-level, transparent and user-friendly approach with common goals (approx. 11%). Other recommendations submitted are to facilitate administrative processes on EU level, to adopt best practices from initiatives such as CLARIN[7] and META-NET, to enforce project evaluation processes and to establish business models and commercialisation plans to raise awareness for the ongoing work and the field of Language Technology in general.

4.3. Part 3: Talent Generation and Retention

The last part of the survey addresses another challenge Europe's LT community is currently facing, i. e., the constantly increasing brain drain (STOA, 2017). Q26 and Q27 assess the incentives needed for early stage researchers to stay in Europe as well as the skills that are mostly demanded in Language Technology and related fields. In order to best address the skill gap, 74% out of all respondents envision closer collaboration between academia and industry (e. g., through job fairs and hackathons). A large percentage of 62% also sees opportunities in the reorganisation of university curriculums, 43% emphasise the importance of fostering a more enterpreneurial culture through specialised course modules, accelerator programmes etc. As for relevant skills the highest priority from both senior as well junior researchers are with around 11% linguistic expertise which compasses all disciplines, followed by strong programming skills (approx. 10%) and expertise in Machine Learning (approx. 8%).

[7]https://www.clarin.eu

5. Conclusions

The survey has shown that there is a profound common interest and passion not only with regard to Multilingual Europe but also in making the ambitious idea of a large-scale, long-term Human Language Project a reality. The answers emphasise that raising awareness for the Language Technology potential in Europe on a political level is more important now than ever before. The upcoming Brexit and trend of highly qualified researchers emigrating to the US leaves the European Language Technology community in a place where change is needed in order to compete with innovative systems and technologies built and research results produced in the US and elsewhere. With regard to opportunities for research and technology development the three most prominent areas to focus on are Machine Translation, Educational and Language Learning technologies as well as Deep Learning and Natural Language Understanding.
On top of that, the survey inspired plenty of positive comments, for example:

- *"This inspired my brains a lot. Thanks for good questions. I think this is the BEST questionnaire I have ever filled! Good luck with your work! Do not hesitate to contact me if you like to ask or discuss more. I would enjoy continuing in this kind of way, it makes me excited!"*
- *"Human Language Project is an excellent initiative."*
- *"Best wishes to the survey – this is one of the most important topics for Europe at the present time."*

Should this submission be accepted for LREC 2018, the final paper will include a more detailed analysis of the survey. If the LREC 2018 Programme Committee wants to organise a dedicated session around more strategic research funding oriented topics, the authors would be happy to contribute the results of the survey, as described in this contribution.

Acknowledgements

CRACKER has received funding from the EU's Horizon 2020 research and innovation programme through the contract CRACKER (grant agreement no.: 645357).

Bibliographical References

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., and Turchi, M. (2017). Findings of the 2017 Conference on Machine Translation (WMT17).

In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark, September. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Georg Rehm et al., editors. (2012). *META-NET White Paper Series: Europe's Languages in the Digital Age*. Springer, Heidelberg, New York, Dordrecht, London. 31 volumes on 30 European languages. http://www.meta-net.eu/whitepapers.

Rehm, G. and Uszkoreit, H. (2013). *META-NET Strategic Research Agenda for Multilingual Europe 2020*. Springer Publishing Company, Incorporated.

Rehm, G., Uszkoreit, H., Dagan, I., Goetcherian, V., Dogan, M. U., Mermer, C., Váradi, T., Kirchmeier-Andersen, S., Stickel, G., Jones, M. P., Oeter, S., and Gramstad, S. (2014). An Update and Extension of the META-NET Study "Europe's Languages in the Digital Age". In *Proceedings of the Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era (CCURL 2014)*, Reykjavik, Iceland, May.

Rehm, G., Hajic, J., van Genabith, J., and Vasiļjevs, A. (2016a). Fostering the Next Generation of European Language Technology: Recent Developments – Emerging Initiatives – Challenges and Opportunities. In Nicoletta Calzolari, et al., editors, *Proceedings of the 10th Language Resources and Evaluation Conference (LREC 2016)*, pages 1586–1592, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Rehm, G., Uszkoreit, H., Ananiadou, S., Bel, N., Bielevičienė, A., Borin, L., Branco, A., Budin, G., Calzolari, N., Daelemans, W., Garabík, R., Grobelnik, M., García-Mateo, C., van Genabith, J., Hajič, J., Hernáez, I., Judge, J., Koeva, S., Krek, S., Krstev, C., Lindén, K., Magnini, B., Mariani, J., McNaught, J., Melero, M., Monachini, M., Moreno, A., Odjik, J., Ogrodniczuk, M., Pęzik, P., Piperidis, S., Przepiórkowski, A., Rögnvaldsson, E., Rosner, M., Pedersen, B. S., Skadiņa, I., Smedt, K. D., Tadić, M., Thompson, P., Tufiş, D., Váradi, T., Vasiļjevs, A., Vider, K., and Zabarskaite, J. (2016b). The Strategic Impact of META-NET on the Regional, National and International Level. *Language Resources and Evaluation*. 10.1007/s10579-015-9333-4.

Rehm, G. (2015). Strategic Agenda for the Multilingual Digital Single Market – Technologies for Overcoming Language Barriers towards a truly integrated European Online Market, April. Version 0.5. April 22, 2015. Prepared by the EU-funded projects CRACKER and LT_Observatory.

Rehm, G. (2016). Language as a Data Type and Key Challenge for Big Data. Strategic Research and Innovation Agenda for the Multilingual Digital Single Market. Enabling the Multilingual Digital Single Market through technologies for translating, analysing, processing and curating natural language content, July. Version 0.9. July 04, 2016. Prepared by the Cracking the Language Barrier federation, supported by the EU-funded projects CRACKER and LT_Observatory.

Rehm, G. (2017). Language Technologies for Multilingual Europe: Towards a Human Language Project. Strategic Research and Innovation Agenda, November. Working title. Version 1.0. To be unveiled at META-FORUM 2017 in Brussels, Belgium, on November 13/14, 2017. Prepared by the Cracking the Language Barrier federation, supported by the EU-funded project CRACKER.

STOA. (2017). Language equality in the digital age – Towards a Human Language Project. STOA study (PE 598.621), IP/G/STOA/FWC/2013-001/Lot4/C2, March 2017. Carried out by Iclaves SL (Spain) at the request of the Science and Technology Options Assessment (STOA) Panel, managed by the Scientific Foresight Unit (STOA), within the Directorate-General for Parliamentary Research Services (DG EPRS) of the European Parliament, March. http://www.europarl.europa.eu/stoa/.

A Survey Questions

Below we list all 29 survey questions, divided into three main blocks as well as two closing questions.

A1. Background, Research Interests, Projects

The first 14 questions focus on the demographic background, research interests and projects of the respondents.

- Q1: Personal details
- Q2: What is the name of the organization you work for?
- Q3: What type of organisation do you work for?
- Q4: What is your company's estimated annual revenue in Euro?
- Q5: What is the size of the organisation (total number of employees)?
- Q6: What is your main role in the organisation?
- Q7: What are the day-to-day responsibilities in your role?
- Q8: What are the key research fields, areas and subareas, methods and applications you work on?
- Q9: Which languages do you mainly work with in your research or offer in your products or services?
- Q10: Which languages would you like to include in your research, products or services in addition - but cannot due to a lack of technologies, tools or resources?
- Q11: In which of the following economic sectors do you see high potential for applications, opportunities for commercial growth or promising target markets for your research, products or services?

- Q12: Which of the following Language Technology applications and services for the Multilingual Digital Single Market could be improved through your research?
- Q13: Where do you see crucial gaps in terms of technologies, tools, or resources, especially with regard to specific languages?
- Q14: What is the biggest challenge the European Language Technology community is currently facing?

A2. Visions for a Future Large-Scale Language Technology Programme

Questions 15-25 focus on the vision of a Language Technology Programme (Human Language Project) in the context of Europe's multilingual challenges and gaps.

- Q15: Do you support the idea of setting up a large-scale Human Language Project?
- Q16: Are there any specific reasons why you do not support the setting up of a Human Language Project? Please specify if possible.
- Q17: Do you think Deep Natural Language Understanding by 2030 is the right vision and an adequate scientific challenge?
- Q18: Which strategic vision would you suggest instead?
- Q19: How long do you think the HLP needs to be so that it can reach the suggested scientific vision and have a significant impact?
- Q20: In the context of a HLP, what are, in your opinion, the (up to) five key challenges Europe needs to work in with regard to: a) Basic research, b) Applied research, c) Innovation, d) Industries/Sectors/Verticals
- Q21: Which are the top three research, technology development, or socio-economic opportunities that you personally envisage the HLP to bring about or to successfully address?
- Q22: Do you have any other additional suggestions or recommendations with regard to the HLP? For example, how it should be organised in terms of priority setting and governance or with regard to strategic guidance
- Q23: How should the Human Language Project be funded?
- Q24: What are, in your opinion, the five key topics,applications, services that must be included in such a platform?
- Q25: Do you have any additional recommendations regarding the setup of the European Language Technology Data andS ervice Platform? For example, regarding the collaboration between data providers, LT providers and LT consumers?

A3. Part 3: Talent Generation and Retention

Questions 26 and 27 focus on concepts for talent generation and retention in Europe.

- Q26: Which technical or soft skills do you personally consider most important for your specific area/projects?
- Q27: How can the skill gap best be addressed?

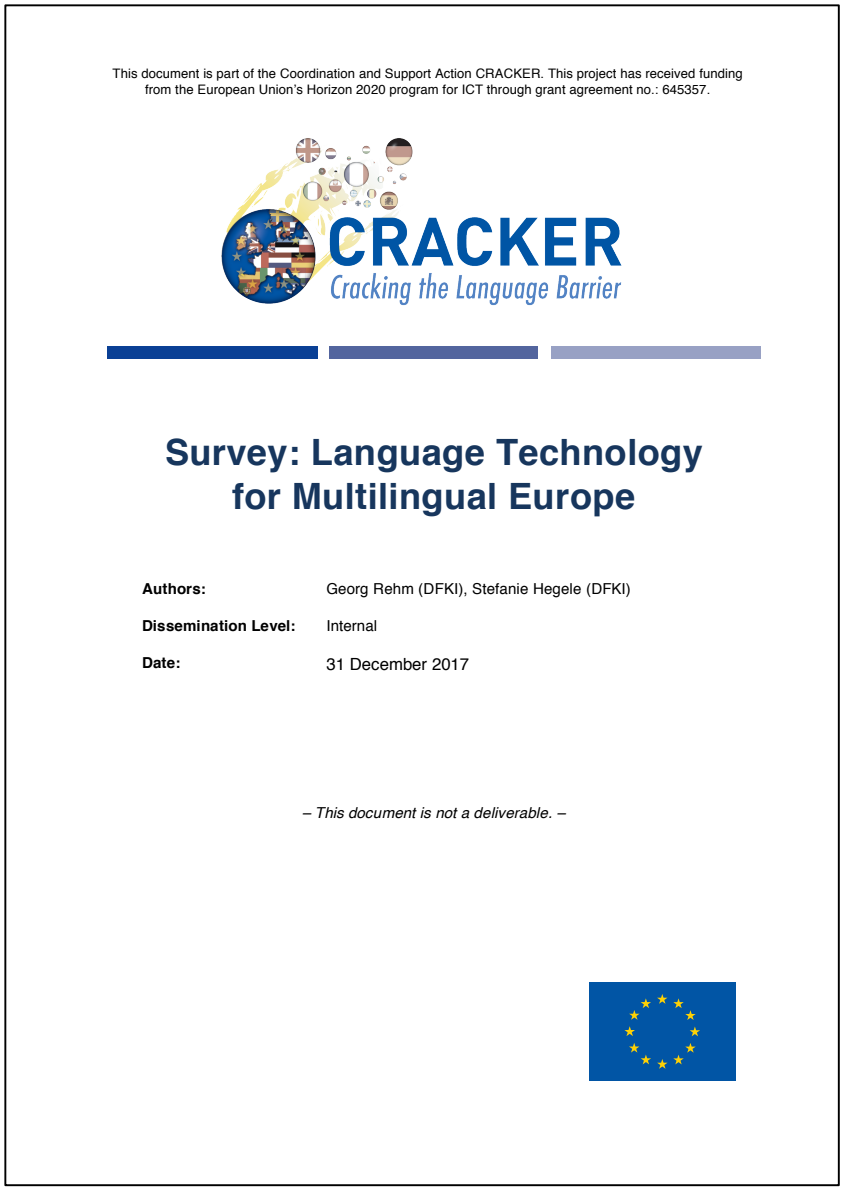
A4. Last but not least

Questions 28 and 29 focus on survey dissemination statistics and final comments.

- Q28: How did you find out about this survey?
- Q29: If you have any additional comments, concerns or suggestions please do not hesitate to share them.

5 Report on the Survey: Language Technology for Multilingual Europe

In the following, we include the full report on the survey “Language Technology for Multilingual Europe”. The included document is not a deliverable. It is meant to be a complement to CRACKER Deliverable D4.3 “Survey on the state of HQMT in industry and LSPs”, as presented and discussed in the first CRACKER review.


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.

CRACKER
Cracking the Language Barrier

Survey: Language Technology for Multilingual Europe

Authors: Georg Rehm (DFKI), Stefanie Hegele (DFKI)

Dissemination Level: Internal

Date: 31 December 2017

– This document is not a deliverable. –

Figure 3: Title page of the Report “Survey: Language Technology for Multilingual Europe”

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.

Survey: Language Technology for Multilingual Europe

Authors: Georg Rehm (DFKI), Stefanie Hegele (DFKI)

Dissemination Level: Internal

Date: 31 December 2017

– This document is not a deliverable. –

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	Internal
Contractual date of delivery	–
Actual date of delivery	–
Deliverable number	This document is not a deliverable.
Deliverable title	Survey: Language Technology for Multilingual Europe
Type	–
Status and version	–
Number of pages	86
Contributing partners	DFKI
WP leader	–
Task leader	–
Authors	Georg Rehm (DFKI), Stefanie Hegele (DFKI)
Internal reviewers	–
EC project officer	Pierre-Paul Sondag (M01-M18), Susan Fraser (M19-M36)
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 Survey Overview 11

1.1 Conceptualization 11

1.2 Dissemination 11

2 Analysis overview 15

2.1 Survey scope 15

2.2 Survey Demographics 15

3 Survey Questions 17

3.1 Personal details 17

3.1.1 How many years of work experience do you have? 17

3.1.2 In which country are you based 18

3.2 What is the name of the organization you work for? 18

3.3 What type of organisation do you work for? 19

3.4 What is your company's estimated annual revenue in Euro? 19

3.5 What is the size of the organisation (total number of employees)? 19

3.6 What is your main role in the organisation? 21

3.7 What are the day-to-day responsibilities in your role? 21

3.8 What are the key research fields, areas and sub-areas, methods and applications you work on? 22

3.8.1 Fields 22

3.8.2 Areas and sub-areas 22

3.8.3 Methods 23

3.8.4 Applications 23

3.9 Which languages do you mainly work with in your research or offer in your products or services? 24

3.10 Which languages would you like to include in your research, products or services in addition – but cannot due to a lack of technologies, tools or resources? 27

3.11 In which of the following economic sectors do you see high potential for applications, opportunities for commercial growth or promising target markets for your research, products or services? 29

3.12 Which of the following Language Technology applications and services for the Multilingual Digital Single Market could be improved through your research? 30

3.13 Where do you see crucial gaps in terms of technologies, tools, or resources, especially with regard to specific languages? 31

3.14 What is the biggest challenge the European Language Technology community is currently facing? 31

3.15 Do you support the idea of setting up a large-scale Human Language Project? 33

3.16 Are there any specific reasons why you do not support the setting up of a Human Language Project? Please specify if possible. 33

3.17 The above-mentioned study suggests, in terms of the HLP's key strategic vision, to concentrate on achieving Deep Natural Language Understanding by 2030. Do you think this is the right vision and an adequate scientific challenge? 33

3.18 Which strategic vision would you suggest instead? 34

3.19 How long do you think the HLP needs to be so that it can reach the suggested scientific vision and have a significant impact? 34

3.20 Let's assume that we have a comfortably funded Human Language Project with a timespan of ca. 5-15 years. What are, in your opinion, the (up to) five key challenges Europe needs to work in with regard to: 35

3.20.1 Basic research 35

3.20.2 Applied research 35

3.20.3 Innovation 36

3.20.4 Industries/Sectors/Verticals 37

3.21 Which are the top three research, technology development, or socio-economic opportunities that you personally envisage the HLP to bring about or to successfully address? 37

3.22 Do you have any other additional suggestions or recommendations with regard to the HLP? For example, how it should be organised in terms of priority setting and governance or with regard to strategic guidance? 39

3.23 How should the Human Language Project be funded? 40

3.24 What are, in your opinion, the five key topics, applications, services that must be included in such a platform? 40

3.25 Do you have any additional recommendations regarding the setup of the European Language Technology Data and Service Platform? For example, regarding the collaboration between data providers, LT providers and LT consumers? 41

3.26 Which technical or soft skills do you personally consider most important for your specific area/projects? 42

3.27 How can the skill gap best be addressed? 42

3.28 Last but not least... 43

3.28.1 How did you find out about this survey? 43

3.28.2 If you have any additional comments, concerns or suggestions please do not hesitate to share them. 43

4 Appendix I – Detailed analysis of open question answers 44

4.1 Questionnaire answer rates 44

4.2 Where do you see crucial gaps in terms of technologies, tools, or resources, especially with regard to specific languages? 46

4.3 What is the biggest challenge the European Language Technology community is currently facing? 48

4.4 Are there any specific reasons why you do not support the setting up of a Human Language Project? Please specify if possible. 50

4.5 Which strategic vision would you suggest instead? 51

4.6 Let's assume that we have a comfortably funded Human Language Project with a timespan of ca. 5-15 years. What are, in your opinion, the (up to) five key challenges Europe needs to work in with regard to: 52

4.6.1 Basic research 52

4.6.2 Applied research 54

4.6.3 Innovation 57

4.6.4 Industries/Sectors/Verticals 59

4.7 Which are the top three research, technology development, or socio-economic opportunities that you personally envisage the HLP to bring about or to successfully address? 62

4.8 Do you have any other additional suggestions or recommendations with regard to the HLP? For example, how it should be organised in terms of priority setting and governance or with regard to strategic guidance? 65

4.9 What are, in your opinion, the five key topics, applications, services that must be included in such a platform? 66

4.10 Do you have any other additional suggestions or recommendations with regard to the HLP? For example, how it should be organised in terms of priority setting and governance or with regard to strategic guidance? 70

4.11 Which technical or soft skills do you personally consider most important for your specific area/projects? 72

5 Appendix II – Additional tables 75

5.1 In which country are you based? 75

5.2 What is your company's estimated annual revenue in Euro? 78

5.3 What is the size of the organisation (total number of employees)? 78

5.4 Which languages do you mainly work with in your research or offer in your products or services? 79

5.5 Which languages would you like to include in your research, products or services in addition – but cannot due to a lack of technologies, tools or resources? 80

5.6 What are the key research fields, areas and sub-areas, methods and applications you work on? 81

5.6.1 Fields 81

5.6.2 Areas and sub-areas 81

5.6.3 Methods 82

5.6.4 Applications 82

5.7 In which of the following economic sectors do you see high potential for applications, opportunities for commercial growth or promising target markets for your research, products or services? 83

5.8 Which of the following Language Technology applications and services for the Multilingual Digital Single Market could be improved through your research? 84

5.9 If you have any additional comments, concerns or suggestions please do not hesitate to share them. 84

Figures

Figure 1: Number of completed surveys by day 14
Figure 2: Overview survey analysis broken down by device 15
Figure 3: Demographic distribution of survey participants (first 25 countries) 16
Figure 4: How many years of work experience do you have? 17
Figure 5: In which country are you based? 18
Figure 6: What type of organisation do you work for? 19
Figure 7: What is the size of the organisation (total number of employees)? 20
Figure 8: What is your main role in the organisation? 21
Figure 9: What are the day-to-day responsibilities in your role? 22
Figure 10: Fields 22
Figure 11: Areas and sub-areas 23
Figure 12: Methods 23
Figure 13: Applications 24
Figure 14: Which languages do you mainly work with in your research or offer in your products or services? A – M 25
Figure 15: Which languages do you mainly work with in your research or offer in your products or services? N – Z 26
Figure 16: Which languages would you like to include in your research, products or services in addition – but cannot due to a lack of technologies, tools or resources? A – M 27
Figure 17: Which languages would you like to include in your research, products or services in addition – but cannot due to a lack of technologies, tools or resources? N – Z 28
Figure 18: In which of the following economic sectors do you see high potential for applications, opportunities for commercial growth or promising target markets for your research, products or services? 29
Figure 19: Which of the following Language Technology applications and services for the Multilingual Digital Single Market could be improved through your research? 30
Figure 20: Do you support the idea of setting up a large-scale Human Language Project? 33
Figure 21: The above-mentioned study suggests, in terms of the HLP's key strategic vision, to concentrate on achieving Deep Natural Language Understanding by 2030. Do you think this is the right vision and an adequate scientific challenge? 33
Figure 22: How long do you think the HLP needs to be so that it can reach the suggested scientific vision and have a significant impact? 34
Figure 23: How should the Human Language Project be funded? 40
Figure 24: How can the skill gap best be addressed? 42
Figure 25: How did you find out about this survey? 43

Summary

Europe is a multilingual society with 24 official member state languages and many additional unofficial and regional languages as well as languages of minorities, immigrants and important trade partners. Nevertheless, day in and day out, language barriers keep severely hampering the free flow of information, thought, ideas, goods and products through our continent. Powerful multilingual technologies as well as cross-lingual and monolingual language technologies, powered by the latest Artificial Intelligence algorithms, have the potential of overcoming language barriers.

The main purpose of this survey was to collect input, feedback and ideas from the European Language Technology research and innovation community in order to assess the currently most prominent research areas, projects and applications, but more importantly to identify the biggest challenges, obstacles and gaps. The community was encouraged to share concrete suggestions and recommendations on how present challenges can be turned into opportunities in the context of the vision of setting up a Human Language Project, i.e., a large-scale funding programme for European Language Technology research, development, innovation and education.

The survey generated an overwhelming response. A total of 634 questionnaires were submitted by respondents from 52 countries (including 37 European countries and 27 Member States). The fact that more than half of the participants are based at well-known universities and research institutions, holding senior roles in their respective fields, is reflected in the high quality of answers provided. The vast majority of experts work in Language Technology, Computational Linguistics, Linguistics, Artificial Intelligence as well as Computer Science and cover a vast spectrum of research areas, methods and applications.

90% of the survey participants state that their research and the resulting applications are mainly focused around the English language. Still prominent, but not nearly as widely in use are the European languages German, Spanish, French, Italian and Dutch. Thus, fundamental gaps with regard to smaller and minority languages have been arising. One severe result is a serious lack of available domain-specific resources which becomes especially evident and crucial in Machine Translation research. This misbalance and the severe digital extinction threat of smaller languages is exactly where almost half of the participants see the future's biggest challenge. Raising awareness for the Language Technology potential in Europe on a political level has become more important than ever before. Discerning is also the limited funding being made available for low-resource languages (mentioned by 8%). Further challenges are posed by cumbersome copyright restrictions for certain data repositories. To overcome these hurdles, better interoperability and standardisation methods need to be established.

On a technical level, other obstacles include the insufficient quality of available natural language processing tools and lacking multilingual technologies in areas like Semantics, Pragmatics, Discourse Analysis, Natural

Language Understanding and Deep Learning (agreed on by 19% of the respondents).
In this context, 97% of all survey participants support the idea of setting up the Human Language Project as a large-scale European funding programme. Moreover, 87% believe that “Deep Natural Language Understanding by 2030” is indeed an adequate scientific challenge the continent should tackle under the umbrella of the Human Language Project. In terms of duration, around 70% suggest a time span of 10-15 years as necessary in order to arrive at satisfactory and sustainable results. Almost 90% agree that the European Commission should be the driving force and main stakeholder when it comes to funding. However, over 50% see also shared responsibilities from the industry and member states as essential for success.
With a sufficiently funded Human Language Project, basic research should have its primary focus on improving current resources and making new high-quality ones publicly available. As the main research topics respondents mentioned Deep Learning and Natural Language Understanding. In addition, there is a need to improve common applications such as Question Answering, Summarisation and Machine Translation. Applied research should mainly be concerned with the investigation of new approaches to Machine Translation and, with regard to supporting human translators, existing CAT tools, especially for a more sophisticated handling of semantic and pragmatic knowledge, as well as the collection of multilingual resources. Of equal importance is the creation of a truly standardized and interoperable open-source platform for data resources tailored to Europe’s needs.
When it comes to innovation the fostering of multilingual and inter-cultural software systems including all languages is regarded as the top priority. This also presupposes better and stronger relations between academia and industry to help foster an innovative start-up culture and in general to seize more business opportunities. For industry, verticals and sectors the most mentioned and therefore key categories are Health and Education.
Identified as the sectors with the highest potential contribution to commercial growth are Education (71%), Information & Communication Technologies (64%) as well as Human Health & Social Work (45%). Specific services and applications that could benefit the Multilingual Digital Single Market comprise better Language Resources and Technologies (73%), Translation Services (46%), Multilingual Solutions for E-Learning (41%) and E-Health (38%).
With regard to Europe’s future and opportunities for research and technology development some of the most promising areas are considered to be Machine Translation, Deep Learning and Natural Language Understanding. Socio-economic opportunities are likely to emerge by guaranteeing better access to multilingual data and providing services for all people. This would also establish a solid basis for the inclusiveness of minorities and people with special needs. Most mentioned domains for applications besides education are health care and e-commerce.

In the face of all these challenges, true multilingualism, realised through powerful language technologies, is a solution to help remove language barriers, to foster collaboration and to create more cultural awareness.

1 Survey Overview

1.1 Conceptualization

This survey was specifically targeted at the European Language Technology Research and Innovation community with the aim to collect valuable input on current research activities, trends as well as visions for the future.
The findings of this survey will serve as an important contribution to the final Strategic Research and Innovation Agenda which is to be officially published at the META-FORUM conference on November 13/14, 2017.
The survey contains 29 questions of which 16 are open questions with free text answers. The remaining ones are a mixture of multiple choice and yes/no questions.
Further, the survey is divided into three main parts covering:

1. Background, research interests and projects
2. Visions for a future large-scale European LT Project
3. Ideas on talent generation and retention in Europe

This division allowed to capture an overview of current and on-going research activities and developments in the field in the first part, reaching early-stage as well as more senior community members. The second part was intended to gather more expert knowledge with regard to visions and concrete plans for future work, in particular steps and prerequisites needed for initializing a large-scale Human Language Technology Project tailored especially to Europe's demands and current opportunities. The third and final part addresses the current challenge of the brain drain the European LT (and also AI) community is experiencing.
The survey underwent two consultation rounds (26th April until 10th May and 10th May until 16th May) and was reviewed by several members of META-NET and the Cracking the Language Barrier federation as well as the responsible project officer from the European Commission.
The survey was designed and set up using Typeform[1], a software that specializes in online form building.

1.2 Dissemination

The survey was officially launched on May 16th and closed on July 4th 2017. As a special incentive to increase the number of participants, those who submitted the survey had the chance to win an iPad.
It was first shared only within a smaller circle of approximately 50 people (mainly the META-NET and Cracking the Language Barrier network) with an appeal to share the survey within their networks and on social media. In a second round a wider audience of more than 4000 people was targeted,

[1] https://www.typeform.com

including participants of former META-FORUM and other conferences as well as respondents of the META-NET Open Letter campaign.

The dissemination timeline looks as follows:

16th May 2017 (Tue):

Official launch

Dissemination through META-NET and Cracking the Language Barrier Board members

Posts on social media (LinkedIn and Facebook) and mailing lists (e.g., Corpora List etc.)

16/17th May 2017 (Tue/Wed):

Emails to Cracking the Language Barrier federation

19th June 2017 (Mon):

Personalised email to META-FORUM and META-NET open letter campaign participants

20th June 2017 (Tue):

Personalised email to members of META

23rd June 2017(Tue):

Email reminder to members of META

27th June 2017 (Tue):

Personalised email reminder to META-FORUM and open letter campaign participants

4th July 2017 (Tue):

Survey deadline

The graph below (Figure 1) shows the total number of survey responses (634) collected over the period of six weeks. Peaks (e.g., on June 19th and June 27th) correspond to our dissemination efforts outlined above.

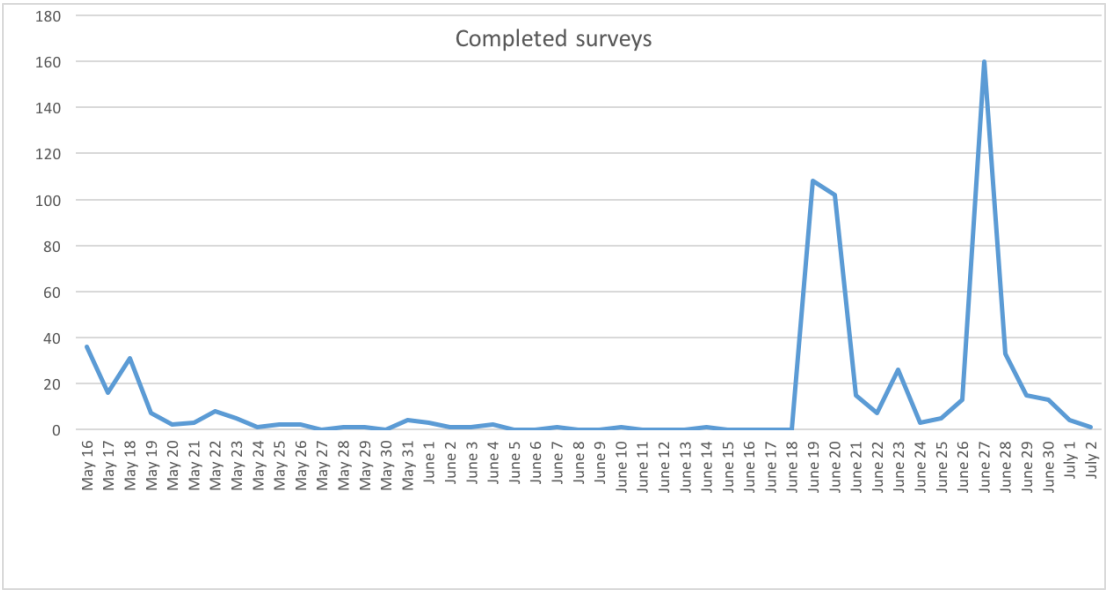


Figure 1: Number of completed surveys by day

2 Analysis overview

This section provides the main statistics regarding scope, distribution and demographics of the survey.
Survey participants were not obliged to answer all questions, but encouraged to contemplate and fill in the ones they feel comfortable with. Chapter 4 displays in detail the statistics for the multiple choice and yes/no questions and qualitative analyses for the open answer questions.

2.1 Survey scope

The survey created an overwhelming feedback with a total of 634 responses and an overall completion rate of 28%; based on previous experience this number must be considered very high. The average time of 35,48 minutes needed for completing the survey indicatives the high-standard and exhaustive quality of the questions answered. Both the completion rate and also the average time needed to fill in the survey demonstrate that the respondents are very passionate about the language topic and about Europe's multilingualism.
For an actual breakdown of the number of completed surveys with average completion times by device see figure 2 below.

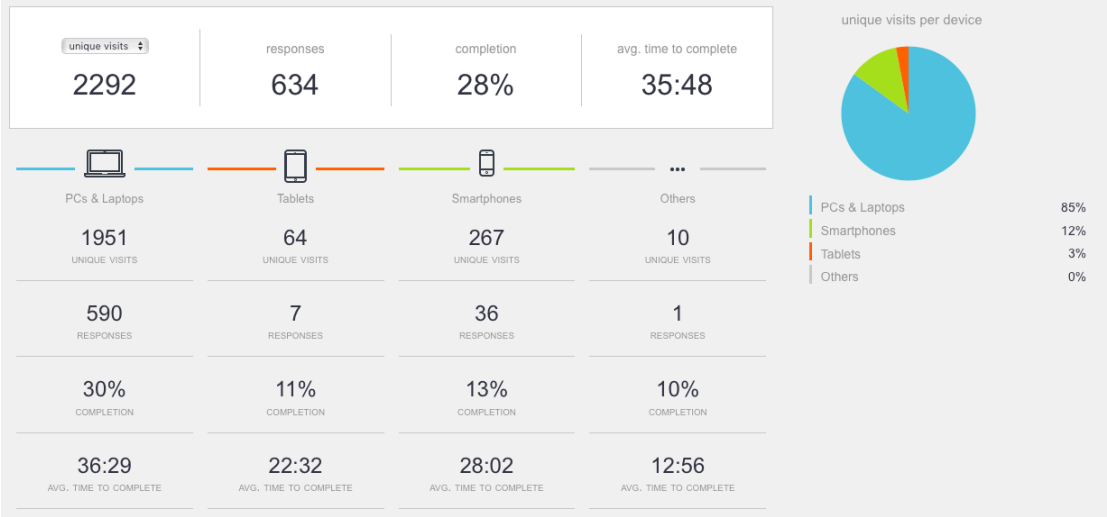


Figure 2: Overview survey analysis broken down by device

2.2 Survey Demographics

One major goal of this survey was to bring the European LT community together and hence reach a wide and demographically distributed audience. Statistics tracked with Google Analytics reveal that the survey was opened by participants in 67 different countries with most views from 1) Germany, 2) Spain, 3) United Kingdom, 4) Italy and 5) United States (see figure 3 below).

A detailed country breakdown of the first 25 participating countries can be found in figure 3 below. A full list of all 52 countries can be found in the appendix (see: In which country are you based?)

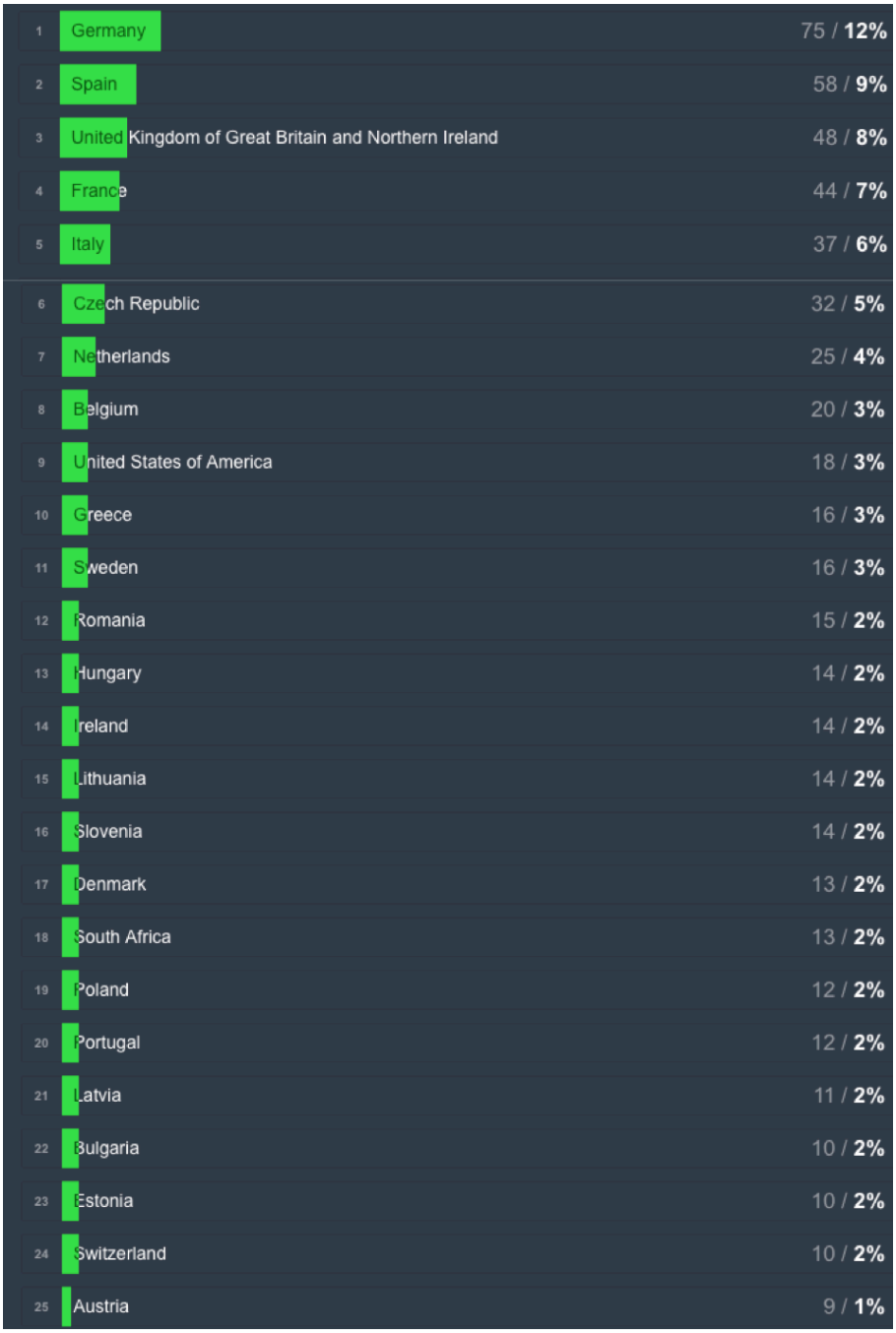


Figure 3: Demographic distribution of survey participants (first 25 countries)

3 Survey Questions

Background, research interests and projects

The first part of the survey consists of 14 questions aiming to collect background information of participants' organisations and their size (and also revenue if applicable) as well as the type of role and day-to-day responsibilities. Further, participants were asked to define the research fields, areas and sub-areas, methods and applications they work on. Particularly important for this survey was to assess in which economic sectors developed applications can be used. Finally, two open answer questions tackle the problems on current gaps (especially with regard to particular languages) and challenges within the European LT community.

3.1 Personal details

In the personal details section, we first asked for name, surname as well as email address of the participants. This information will be kept confidential and won't be attached to this report. The two other sub-questions were:

3.1.1 How many years of work experience do you have?

634 out of 634 participants (100%) answered this question.

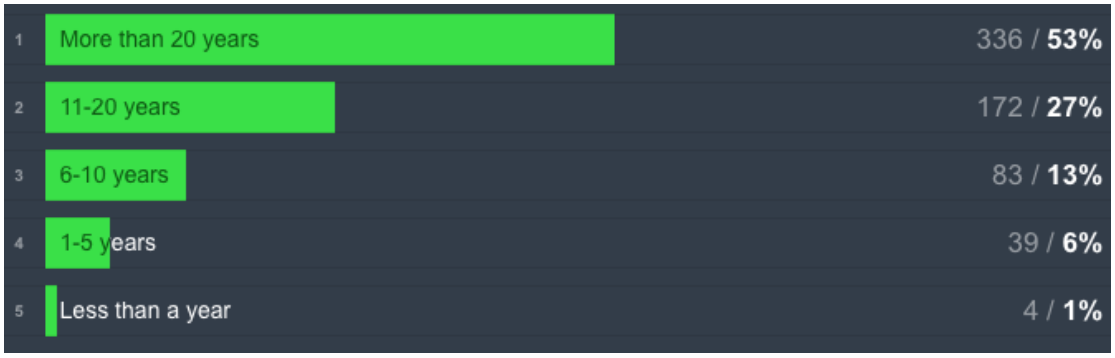


Figure 4: How many years of work experience do you have?

3.1.2 In which country are you based

631 out of 634 participants (approx. 100%) answered this question.

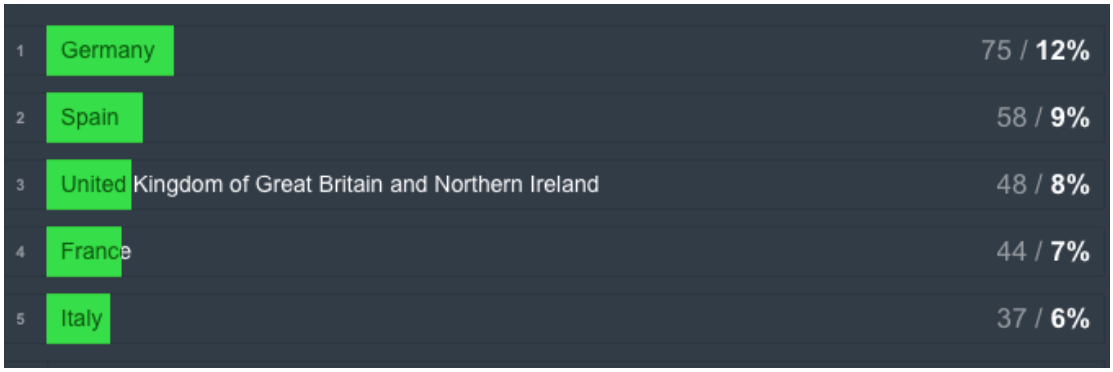


Figure 5: In which country are you based?

A complete country list can be found in the appendix (see: In which country are you based?).

3.2 What is the name of the organization you work for?

613 out of 634 participants (approx. 97%) answered this question.

Since this was an open answer question and participants could freely choose the format to enter their organisation, a detailed analysis of this question was not conducted.
Among the most frequently mentioned organisations were: Charles University in Prague, Vilnius University, University of Copenhagen, DFKI, University of Tartu, University of Edinburgh, CNRS, Ghent University and Bangor University.

3.3 What type of organisation do you work for?

634 out of 634 participants (100%) answered this question.

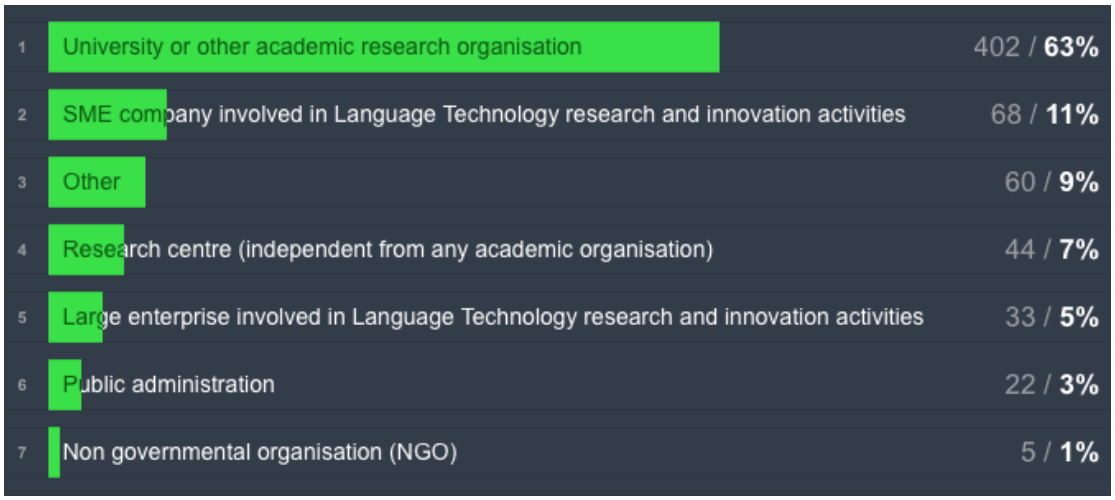


Figure 6: What type of organisation do you work for?

3.4 What is your company's estimated annual revenue in Euro?

60 out of 634 participants (approx. 10%) answered this question.

Since this was an open answer question and participants could freely choose the format to enter their annual revenue. All of the 60 participants who answered this question are affiliated with either large enterprises (18 out of 60) or SMEs (42). As for the large enterprises we have provided revenue numbers between 3.000.000 and 77.000.000.000.
For the SMEs the same analysis did not prove to be reliable. Since this was an open text field we collected low numbers such as “3” where it is not entirely sure if the person really meant 3 euros or eventually 3k (meaning 3000 euros).
Hence, a detailed analysis for the revenue numbers was only done for the large enterprises and can be found in the appendix (see: What is your company's estimated annual revenue in Euro?)

3.5 What is the size of the organisation (total number of employees)?

631 out of 634 participants (approx. 100%) answered this question.

An overview of this question can be found in figure 7 below. A further analysis shows that participants based at universities or research centres opted for giving the total number of employees at their institution rather than their specific research department. As a result, we do not have actual numbers about the size of the research labs.

However, we have more information about large enterprises and SMEs. As outlined in figure 6 (see previous question 4.3) 33 participants are affiliated with large enterprises (among the most popular ones with more than 10000 employees are: Microsoft, IBM, Oracle, Intel and Nuance). A full list of all companies presented in this survey can be found in the appendix (see: What is the size of the organisation (total number of employees)?

Figure 7: What is the size of the organisation (total number of employees)?

3.6 What is your main role in the organisation?

634 out of 634 participants (100%) answered this question.

The distribution shows that a large majority of participants hold senior roles at their respective organisations (e.g. professor, senior researcher, research group leader) which is reflected in the high quality of answers collected. This information about the job roles seen in context with the seniority level (question 4.1 displays that 53% have more than 20 years of work experience) and the demographic distribution of 52 countries clearly leads to the conclusion that the survey represents a wide and high-level range of the European Language Technology research and innovation community.

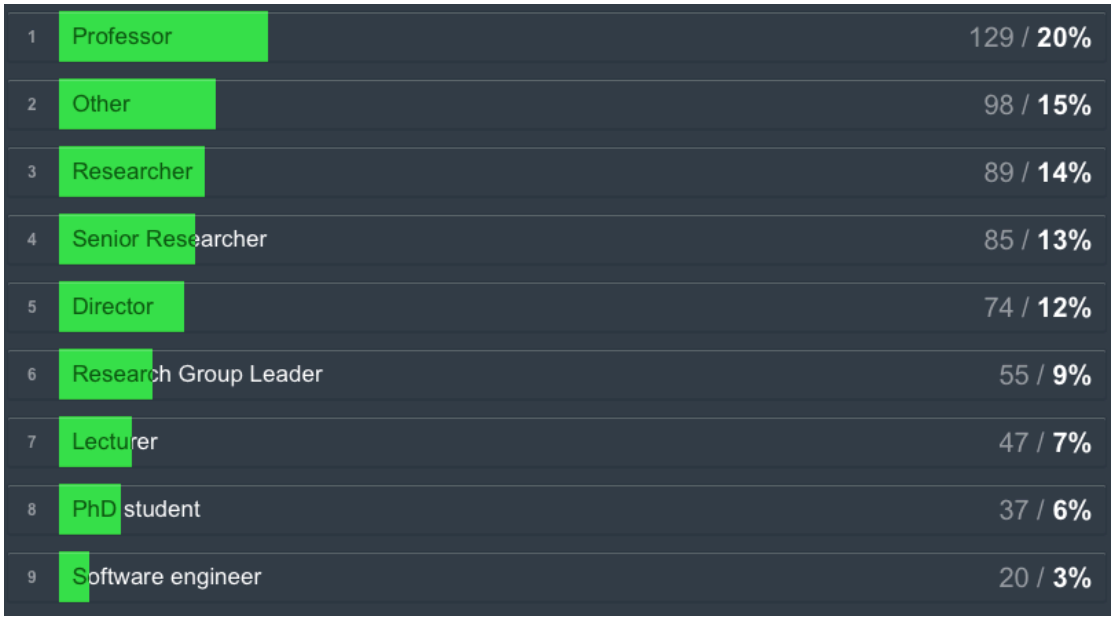


Figure 8: What is your main role in the organisation?

3.7 What are the day-to-day responsibilities in your role?

623 out of 634 participants (approx. 98%) answered this question.

When it comes to day-to-day responsibilities 71% of all participants state an involvement in research, closely followed by 52% naming project management and 43% project execution as their most crucial tasks. This variety of engagement and responsibilities allows to get insightful input on concrete research topics (for basic and applied research as well as innovation topics), methods and best practises as asked in the second part of the survey. In addition, the vast expertise in management and project acquisition indicates competence for answering questions related to strategic planning as well as questions requiring a wider perspective on the field such as the impact Language Technology could have on the Digital Single Market.

Figure 9: What are the day-to-day responsibilities in your role?

3.8 What are the key research fields, areas and sub-areas, methods and applications you work on?

3.8.1 Fields

625 out of 634 participants (approx. 99%) answered this question.

The most prominent fields are Language Technology (64%) and Computational Linguistics (56%) which includes the exact target group the survey was supposed to reach. With the growing importance of Artificial Intelligence for many applications, having 39% with expertise in this field can lead to insightful answers especially with regard to research future topics.

A complete list of all different fields can be found in the appendix (see: In which country are you based?).

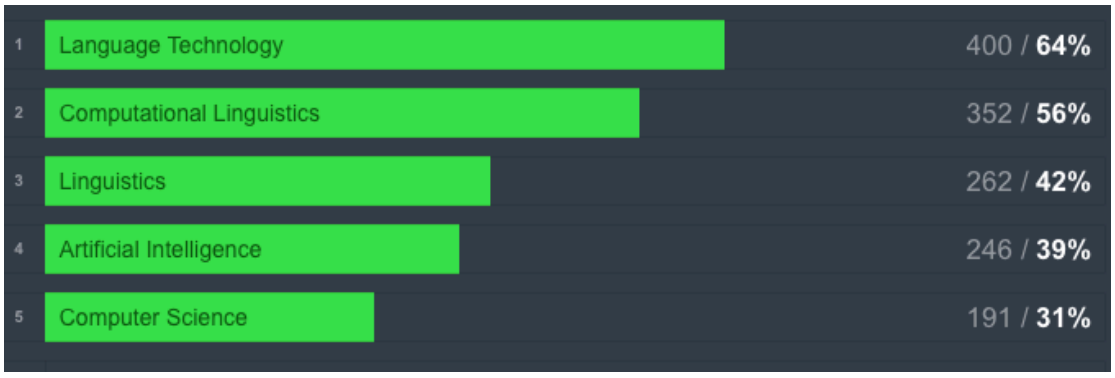


Figure 10: Fields

3.8.2 Areas and sub-areas

618 out of 634 participants (approx. 97%) answered this question.

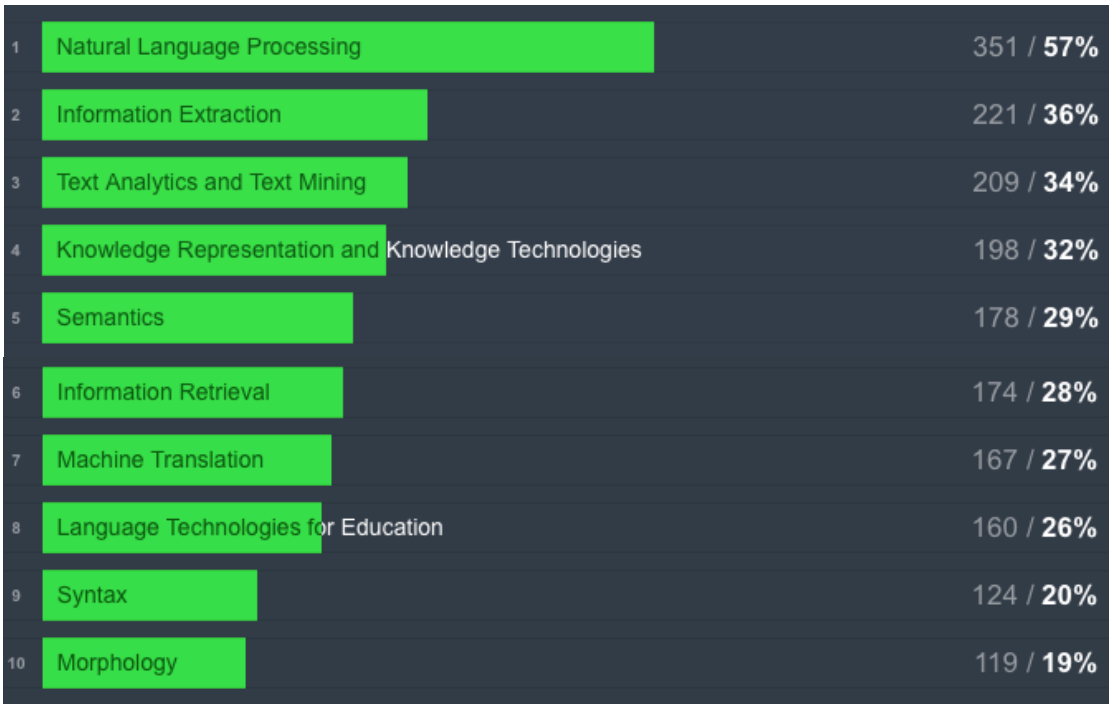


Figure 11: Areas and sub-areas

A complete list of all different areas and sub-areas can be found in the appendix (see: Areas and sub-areas).

3.8.3 Methods

596 out of 634 participants (approx. 94%) answered this question.

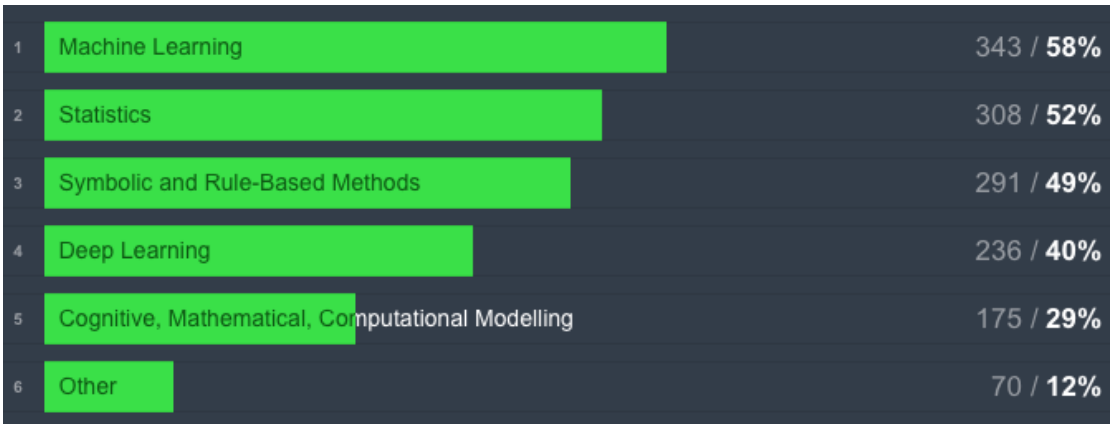


Figure 12: Methods

A complete list of all methods can be found in the appendix (see: Methods).

3.8.4 Applications

616 out of 634 participants (approx. 97%) answered this question.

Figure 13: Applications

3.9 Which languages do you mainly work with in your research or offer in your products or services?

598 out of 634 participants (approx. 94%) answered this question (Lang. A-M).
409 out of 634 participants (approx. 65%) answered this question (Lang. N-Z).

The results unequivocally show that English (90%) is the main language used for research. Most people marked at least 2 languages with English usually being one of them. Frequently used, though not even half as popular, are the big European languages: Spanish (49%) German (41%), French (37%) and Italian (23%). This reflects to a large extend the findings of the META-NET White papers which emphasise the thread of a digital extinction for most European languages if no appropriate measures are taken in the near future.[2]

Figure 14 below lists languages A-M and Figure 15 N-Z. A detailed list of all languages can be found in the appendix (see: Which languages do you mainly work with in your research or offer in your products or services?)

[2] http://www.meta-net.eu/whitepapers/overview

Figure 14: Which languages do you mainly work with in your research or offer in your products or services? A – M

Figure 15: Which languages do you mainly work with in your research or offer in your products or services? N – Z

3.10 Which languages would you like to include in your research, products or services in addition – but cannot due to a lack of technologies, tools or resources?

302 out of 634 participants (approx. 48%) answered this question (Lang. A-M).
279 out of 634 participants (approx. 44%) answered this question (Lang. N-Z).

Figure 16 below lists languages A-M and Figure 17 N-Z. A detailed list of all languages can be found in the appendix (see: Which languages would you like to include in your research, products or services in addition – but cannot due to a lack of technologies, tools or resources?)

Figure 16: Which languages would you like to include in your research, products or services in addition – but cannot due to a lack of technologies, tools, resources? A – M

Figure 17: Which languages would you like to include in your research, products or services in addition – but cannot due to a lack of technologies, tools, resources? N – Z

3.11 In which of the following economic sectors do you see high potential for applications, opportunities for commercial growth or promising target markets for your research, products or services?

619 out of 634 participants (approx. 98%) answered this question.

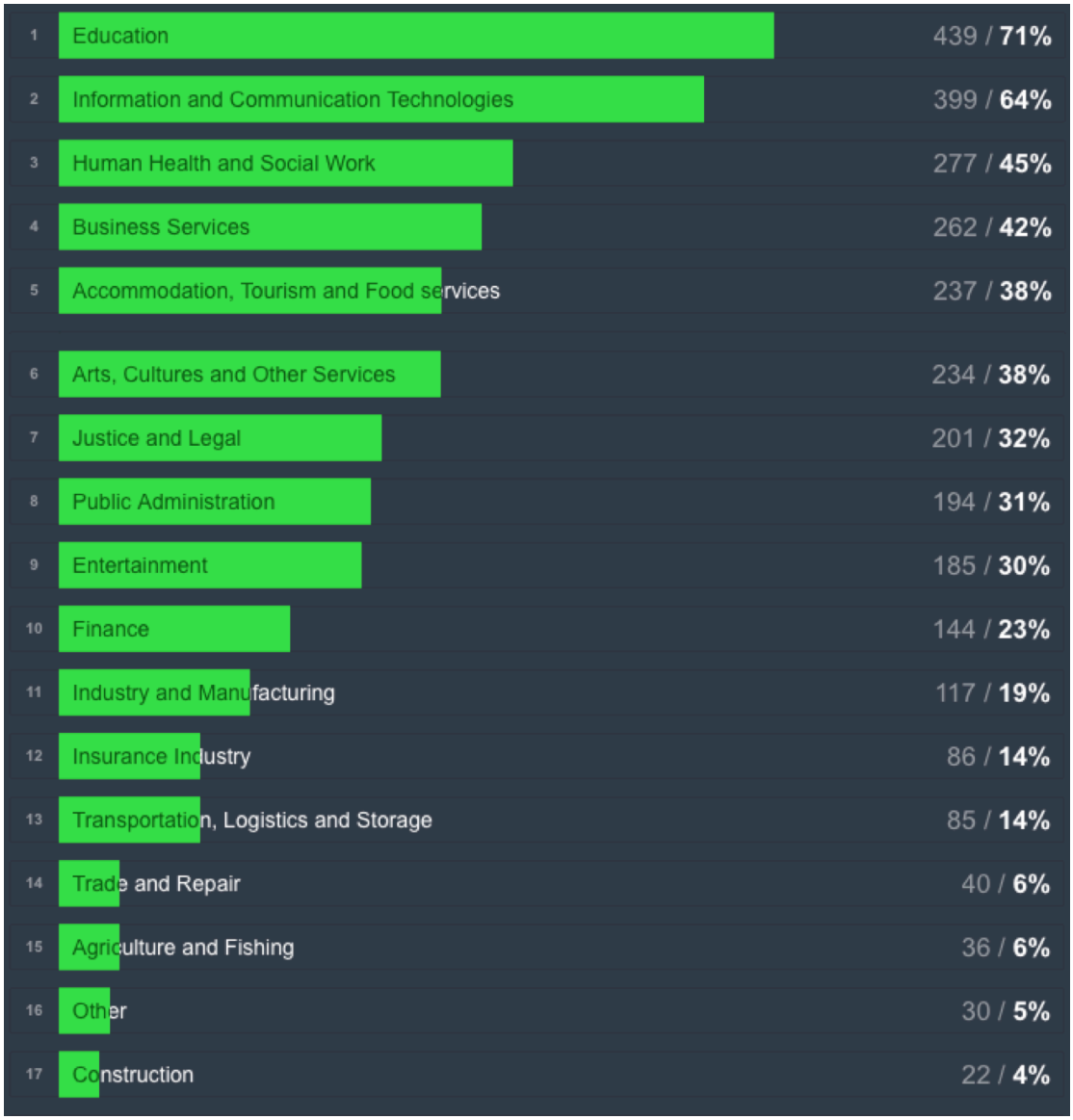


Figure 18: In which of the following economic sectors do you see high potential for applications, opportunities for commercial growth or promising target markets for your research, products or services?

A detailed list of all economic sectors can be found in the appendix (see: In which of the following economic sectors do you see high potential for applications, opportunities for commercial growth or promising target markets for your research, products or services?).

3.12 Which of the following Language Technology applications and services for the Multilingual Digital Single Market could be improved through your research?

605 out of 634 participants (approx. 95%) answered this question.

Figure 19: Which of the following Language Technology applications and services for the Multilingual Digital Single Market could be improved through your research?

A detailed list of all Language Technology applications and services can be found in the appendix (see: Which of the following Language Technology applications and services for the Multilingual Digital Single Market could be improved through your research?).

3.13 Where do you see crucial gaps in terms of technologies, tools, or resources, especially with regard to specific languages?

407 out of 634 participants (approx. 64%) answered this question.

Almost 40% of all participants who answered this question state that there is insufficient research being done for minority languages and dialects, directly resulting in a lack of available resources. This lack becomes most evident in Machine Translation applications for smaller European languages as well as other standard NLP systems which are missing. Also, more domain specific resources are needed.
Problematic is also the limited funding available for low resource languages. Further challenges are posed by copyright restrictions for certain data sets, even when only used for research purposes. More interoperability and standardisations needs to be established as well.
When it comes to research areas the most frequently mentioned are Semantics and Pragmatics, Natural Language Understanding and Processing and speech applications. There is also an interest in educational tools and applications with advanced human-computer interaction, dialogue understanding and better modelling of linguistic knowledge.
Gaps are also existent for the data processing of multimodal input. Another demand from the language service industry is a better and more efficient integration of translation systems since translators still need to use a variety of different stand-alone tools.
A detailed list and more exhaustive summary of all answers can be found in the appendix (see: Where do you see crucial gaps in terms of technologies, tools, or resources, especially with regard to specific languages?).

3.14 What is the biggest challenge the European Language Technology community is currently facing?

408 out of 634 participants (approximately 64%) answered this question.

The biggest challenge the European Language Technology is facing at the moment is the current status and digital extinction threat of smaller languages. Almost a fifth of all provided answers stress the importance of keeping multilingualism in Europe alive. With the dominance of the English language, researchers are often given little incentive to focus on smaller and minority languages. For instance, when it comes to publishing there is a strong bias towards incorporating results for English. The same is true for LT funding being made available for non-English projects.
A consequence of this bias, mentioned by many survey participants, is the lack of available data resources for smaller languages which are especially needed to improve the quality of current Machine Translation systems.
Besides other shortcomings regarding the quality of available NLP tools in general and expertise in research areas like Semantics, Pragmatics,

Discourse Analysis, Natural Language Understanding and Deep Learning, the answers show that raising awareness for the LT potential in Europe on a political level is more important than ever before. The upcoming Brexit and the trend of qualified researchers emigrating to the U.S. leaves the European LT community in a place where change is needed in order to compete with innovative systems and tools build in the U.S. On a political level this involves more commitment from the European Commission as well as the member states. This also includes the removal of bureaucratic hurdles and the establishment of more standardization and easier processes for licensing and access issues.

In addition, an industry-focus on development and use of applications needs to be encouraged. Also, brought up multiple times in this context is the urge for better usability and accessibility of applications for people with no knowledge of English, elderly citizens as well as those with disabilities. This indicates a need for better localisation software for websites and apps as well as stand-alone applications.

A detailed list and more exhaustive summary of all answers can be found in the appendix (see: What is the biggest challenge the European Language Technology community is currently facing?).

Visions for a future large-scale European LT Programme

3.15 Do you support the idea of setting up a large-scale Human Language Project?

623 out of 634 participants (approx. 98%) answered this question.

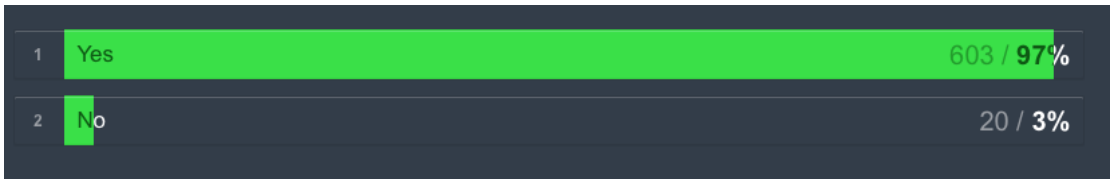


Figure 20: Do you support the idea of setting up a large-scale Human Language Project?

3.16 Are there any specific reasons why you do not support the setting up of a Human Language Project? Please specify if possible.

Out of 634 people only 20 were in disfavour of a large-scale Human Language Project. Reasons given are the previous unsuccessful attempts of similar projects which did not benefit the whole community at the time. Too much bureaucracy, a lack of focus and unclear goals led to a waste of money and resources in the past.
Another point made is the bias being introduced through such research initiatives as they often hinder innovation and hardly ever lead to scientific breakthroughs.
A detailed list and more exhaustive summary of all answers can be found in the appendix (see: Are there any specific reasons why you do not support the setting up of a Human Language Project? Please specify if possible.).

3.17 The above-mentioned study suggests, in terms of the HLP's key strategic vision, to concentrate on achieving Deep Natural Language Understanding by 2030. Do you think this is the right vision and an adequate scientific challenge?

614 out of 634 participants (approx. 97%) answered this question.

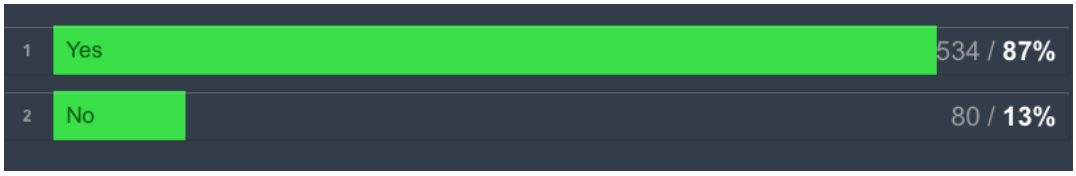


Figure 21: The above-mentioned study suggests, in terms of the HLP's key strategic vision, to concentrate on achieving Deep Natural Language Understanding by 2030. Do you think this is the right vision and an adequate scientific challenge?

3.18 Which strategic vision would you suggest instead?

Alternative strategic visions presented in the survey shift the focus from solely focusing on Deep Natural Understanding since it can hardly solve all challenges the LT field is posing. Those who estimate 2030 as too early for such a big breakthrough advise to prioritize solid basic research and ensure the creation of well-edited language resources first.
Also criticized was the general setup of EU projects and their lack of flexibility and agility. Some even stressed that visions in general are not necessary, but that a better approach would be to support development work done by smaller teams and individuals and also foster international collaboration and education.
Another prevalent opinion was that 2030 as picked deadline is too late. In order to stay competitive scientific breakthroughs need to happen far earlier.
It was also suggested to slightly move away from traditional project management and to create a sort of working marketplace of ideas with appropriate infrastructure (data, computing resources, funding) with smaller milestones.
A detailed list and more exhaustive summary of all answers can be found in the appendix (see: Which strategic vision would you suggest instead?).

3.19 How long do you think the HLP needs to be so that it can reach the suggested scientific vision and have a significant impact?

613 out of 634 participants (approx. 97%) answered this question.

With 35% of people in favour of a run-time of 10-15 years and 24% of more than 15 years there is strong tendency for a HLP set-up of at least 10 years (ideally longer) in order to be impactful.

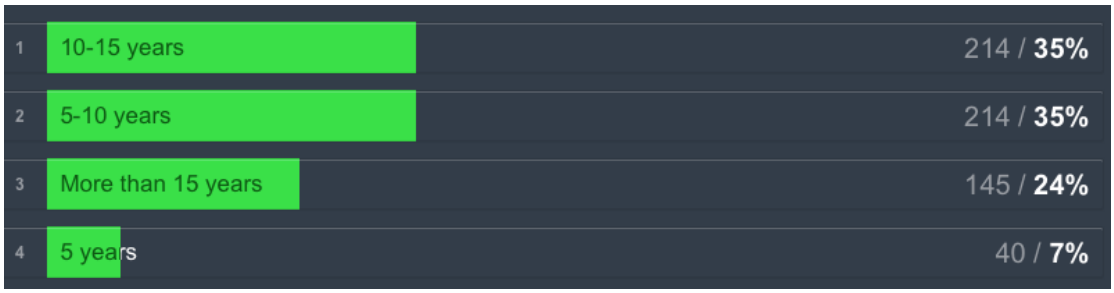


Figure 22: How long do you think the HLP needs to be so that it can reach the suggested scientific vision and have a significant impact?

3.20 Let's assume that we have a comfortably funded Human Language Project with a timespan of ca. 5-15 years. What are, in your opinion, the (up to) five key challenges Europe needs to work in with regard to:

3.20.1 Basic research

422 out of 634 participants (approx. 66,6%) answered this question.

As basic research is concerned a majority mentioned the further development of existing corpora, resources in general, ontologies, dictionaries, terminology repositories, treebanks etc. A focus hereby is to also improve the quality of data annotations. For better and easier accessibility when it comes to copyright issues etc. an effective legal framework should be put in place.
Besides, basic research should be centred around deep learning, neural networks and brain research and Natural Language Understanding. A majority also highlighted the need to further work on existing NLP tasks and tools. Among the most frequently mentioned application areas were: Question Answering, Summarisation, Information extraction and sentiment analysis. Approximately the same number of people saw Machine Translation as the key area for basic research. This includes more work on newer neural approaches, but also a review of existing evaluation and quality assurance methods.
Multilinguality is another significant aspect that needs to be taken into account. Not only from a technical point of view should cross-linguality be incorporated into models, grammar formalisms and NLP applications, but also from a political point of view. The digital extinction threat of many of the smaller and low-resourced languages should be brought to attention on a wider European level.
Also, seen as essential is more research in Linguistics especially in the field of Semantics, but also in Syntax, Morphology, Phonetics and Phonology. Knowledge modelling and meaning was also discussed as important for future work. For speech related research there were mentions for ASR, research on the structure of speech and speech production.
Considered as significant was also research in areas related to Psycholinguistics (language acquisition etc.) and Sociolinguistics covering topics such as the social implication a large-scale Human Language Project would have and the ethics involved.
Finally, it was stressed that collaboration is key for future success. Interdisciplinary work and communication should be strengthened as well as the engagement from political and governmental players.
A detailed list and more exhaustive summary of all answers can be found in the appendix (see: Basic research).

3.20.2 Applied research

343 out of 634 participants (approx. 54%) answered this question.

The predominantly mentioned area for applied research is Machine Translation, covering corpus statistical as well neural techniques. In this context, the need for more advanced computer-assisted translation (CAT) was frequently highlighted.
Seen as almost equally important is the improvement of multilingual resources, databases and terminology repositories, allowing for standardization and interoperability. In addition, there is a need for improved open-source platforms with a wide range of available systems and applications and truly open and unencumbered data and code repositories.
Besides improvements in MT and the availability of language resources the need to continuously raise awareness for the importance of multilingual Language Technology to secure funding is emphasised.
Next to basic NLP and linguistic tools the most mentioned applied research areas are: conversational interfaces and agents, multimodal machine interaction, natural language understanding, dialogue systems and speech technologies.
In order to ensure effective and high quality applied research more time and effort needs to be invested in strong collaboration and communication between academia and industry. Various people suggest that industry should take on the leadership position in this field to ensure better lab to market procedures. Among the other application areas are: ASR, voice recognition, summarisation, knowledge representation and extraction, context aware systems, natural language generation and question answering.
There is also a demand for systems guaranteeing inclusiveness of elderly and people with disabilities and contributing to more social transparency like fake news and hate speech detection.
A detailed list and more exhaustive summary of all answers can be found in the appendix (see: Applied research).

3.20.3 Innovation

257 out of 634 participants (approx. 41%) answered this question.

When it comes to innovation the inclusion of all languages and fostering of inter-cultural systems is mentioned as top priority. This also presupposes better and stronger relations between academia and industry which help foster innovative start-up cultures and help to seize more business opportunities.
As with basic and applied research most answers given underline the significance of improvements in machine translation, especially when it comes to more sophisticated handling of semantic and pragmatic knowledge. Prevalently mentioned are neural and spoken MT application. Besides, the answers show demand and interest for the further development of interactive NLP agents, personal trainers and assistants and dialogue systems.
Also stressed is the need to bring together knowledge and methods developed for different fields and domains, e.g. e-health, e-government, e-

justice and other public domains. All applications and systems should also be made more accessible to a wider and more inclusive audience. Further enhancements for language learning and other educational systems and tools should also be made. In addition, there is also an interest for more advanced visualizations and interfaces, more innovation and new tools incorporating Natural Language Understanding as well as Speech Language Understanding, seamless human-computer and human-robot interactions. Other interests circle around neural networks, deep understanding and brain science, voice and speech recognition and generation, approaches to different multimodal communication approaches and emotion-aware and intention reading applications.
As also indicated in the answers of other questions there is a strong demand for open source frameworks with easy accessibility.
A detailed list and more exhaustive summary of all answers can be found in the appendix (see: Innovation).

3.20.4 Industries/Sectors/Verticals

242 out of 634 participants (approx. 38%) answered this question.

Mostly mentioned in this category is Health. Education comes in second. These two industry sectors are closely followed by Tourism and Travel, Law and Justice, Translation, E-Commerce, Entertainment (incl. arts, creativity, culture and cultural heritage), Media, Business (incl. various services and business intelligence applications), Security, Public services and Administration, Government and Finance.
In the context of industries, sector and verticals the necessity of an on-going knowledge transfer and effective collaboration between academia and industry is once again highlighted.
A detailed list and more exhaustive summary of all answers can be found in the appendix (see: Industries/Sectors/Verticals).

3.21 Which are the top three research, technology development, or socio-economic opportunities that you personally envisage the HLP to bring about or to successfully address?

337 out of 634 participants (approx. 53%) answered this question.

With regard to opportunities for research and technology development the three most prominent areas are Machine Translation, educational and language learning technologies as well as Deep Learning and Natural Language Understanding.
Also suggested as fields where the HLP can have a major impact are: human machine interaction, conversational assistants and chat bots, various NLP tasks where advanced semantics, pragmatics and discourse are needed,

research in human brain processing and cognitive systems as well as social media analysis entailing opinion mining and sentiment analysis.
A detailed list and more exhaustive summary of all answers can be found in the appendix (see Which are the top three research, technology development, or socio-economic opportunities that you personally envisage the HLP to bring about or to successfully address?).

3.22 Do you have any other additional suggestions or recommendations with regard to the HLP? For example, how it should be organised in terms of priority setting and governance or with regard to strategic guidance?

179 out of 634 participants (approx. 28%) answered this question.

A quarter of all suggestions and recommendations stresses the importance of protecting all European languages and thereby increasing intercultural understanding. From a governance point of view this can only be achieved with governments and member states working closely together. This requires that all languages are treated equally, especially when it comes to funding and the development of necessary resources. Such an endeavour builds on a clear and standardized agenda with pre-agreed milestones that guarantee inclusiveness and efficient collaboration of research, industry, economics and politics.
When it comes to strategic guidance many were in favour of funding smaller scale projects, starting bottom-up with smaller goals. This would avoid heavy bureaucracy with big operators and big consortia on EU level. Many also suggested to primarily focus on basic and applied research rather than innovation.
Regarding the governance of an actual Human Language Project, it was suggested to put democratic organization processes in place, e.g. with shifting presidents and elected committee and board members among institutions and countries.
Also frequently mentioned was the need to reposition the strategy of EU research with focus on ideas and scientific breakthroughs in order to diversify from the U.S and large corporation paradigms. This involves fostering strong collaborations between numerous stakeholders, better school and university education with more incentives for young researchers, integration of user and customer experiences, following market driven approaches to ensure industrial growth as well as established feedback processes through regular surveys etc.
A detailed list and more exhaustive summary of all answers can be found in the appendix (see: Do you have any other additional suggestions or recommendations with regard to the HLP? For example, how it should be organised in terms of priority setting and governance or with regard to strategic guidance?).

3.23 How should the Human Language Project be funded?

584 out of 634 participants (approx. 92%) answered this question.

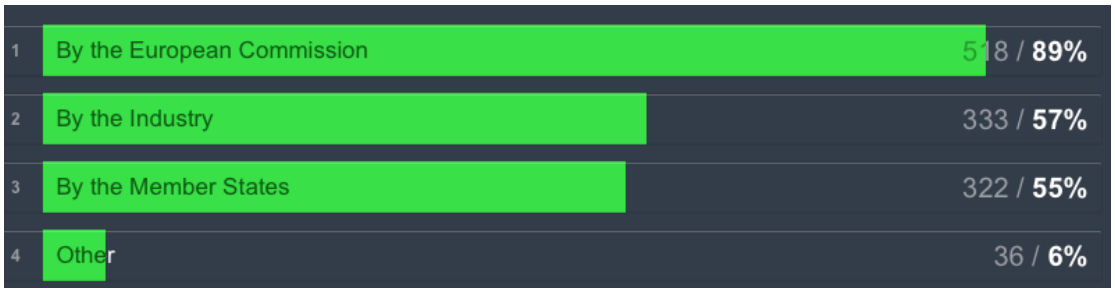


Figure 23: How should the Human Language Project be funded?

3.24 What are, in your opinion, the five key topics, applications, services that must be included in such a platform?

301 out of 634 participants (approx. 47%) answered this question.

Named as the most important application is Machine Translation and related translation memory, CAT and quality assurance services. Considered as almost equally important are the availability of download services for multilingual resources including ontologies, lexicons, dictionaries etc.
Among the topics most relevant for the development of future applications and services are: education, health, e-participation and e-government as well as law and legal services.
As for further applications, a more in-depth development of already existing NLP tools is mentioned and especially speech applications. More specifically, the most frequently mentioned applications are: information extraction and retrieval, summarisation, search systems and intelligent assistants.
Besides, resources should not only be available, but also easily accessible. In addition, it should be facilitated to compile custom-tailored corpora. User-friendly dashboards should also be made available and constantly be improved based on user feedback.
A detailed list and more exhaustive summary of all answers can be found in the appendix (see: What are, in your opinion, the five key topics, applications, services that must be included in such a platform?).

3.25 Do you have any additional recommendations regarding the setup of the European Language Technology Data and Service Platform? For example, regarding the collaboration between data providers, LT providers and LT consumers?

143 out of 634 participants (approx. 23%) answered this question.

A quarter of all answers provided emphasize the importance of easy accessibility and open licensing for available tools and data. Commonly agreed on exchange formats and standards also need to be set up.
When it comes to stakeholder engagement the trend goes towards involvement of all stakeholders, namely data providers, LT providers and LT consumers. Effective communication requires a unified, high-level, transparent and user-friendly approach with common goals. Especially collaboration between research partners in different countries (both within and outside of Europe) should be eased by providing better visa regulations etc. This also aligns with the request for less complicated management and administration process of EU level. Other recommendations are to adopt best practices from former project, to enforce project evaluation processes and to establish business models and commercialization plan to raise awareness for the ongoing work and the field of Language Technology in general.

Talent generation and retention

3.26 Which technical or soft skills do you personally consider most important for your specific area/projects?

454 out of 634 participants (approx. 72%) answered this question.

Mentioned with almost equal frequency are advanced linguistic knowledge and programming skills. Linguistic expertise encompasses hereby all disciplines including semantics, syntax, phonetics, formal linguistics, corpus linguistics etc. For programming languages Python closely followed by Java. R, C++ and Perl were also mentioned.
As necessary technical skills or relevant disciplines people list: machine learning, statistics, mathematics, deep learning and neural networks, computer science, IT as well as computational linguistics.
Considered as the most essential soft skills are collaboration, team work and networking as well as innovative thinking, creativity and proactivity. Also highlighted as fundamental are language skills, curiosity paired with willingness and eagerness to learn alongside project management, analytical and intercultural communication skills.
A detailed list and more exhaustive summary of all answers can be found in the appendix (see: Which technical or soft skills do you personally consider most important for your specific area/projects?).

3.27 How can the skill gap best be addressed?

560 out of 634 participants (approx. 88%) answered this question.

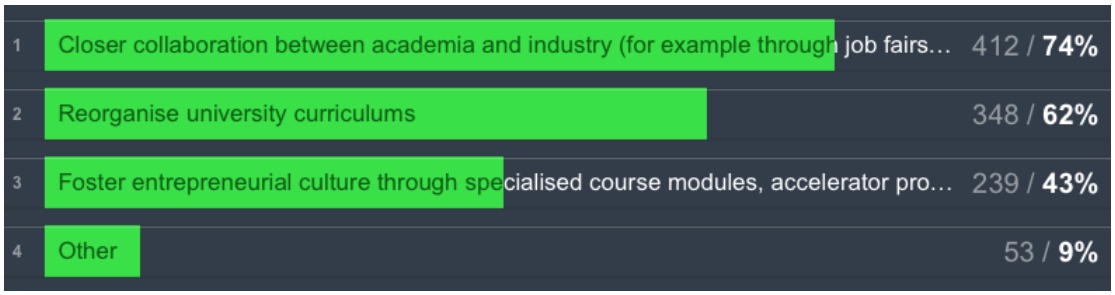


Figure 24: How can the skill gap best be addressed?

3.28 Last but not least...

3.28.1 How did you find out about this survey?

602 out of 634 participants (approx. 95%) answered this question.

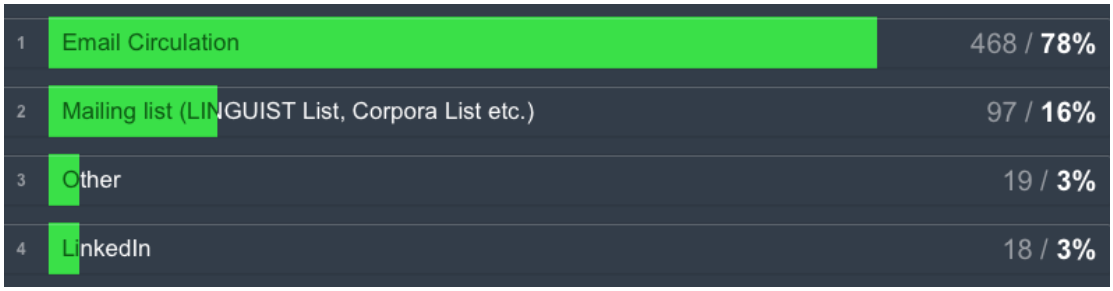


Figure 25: How did you find out about this survey?

3.28.2 If you have any additional comments, concerns or suggestions please do not hesitate to share them.

525 out of 634 participants (approx. 82%) answered this question.

Most of the final comments contained thank you and good luck notes. Also, many participants expressed interest in getting updates on the findings of this survey as well as information on future progress of a large-scale Human Language Project and how to get involved.
Other comments pick up topics from previous questions to repeatedly emphasise the importance of more funding, interdisciplinary, open source approaches, EU Commission and member states engagement, talent retention etc.
A detailed list and more exhaustive summary of all answers can be found in the appendix (see: If you have any additional comments, concerns or suggestions please do not hesitate to share them.).

The overall feedback for this survey has been very positive:

"This inspired my brains a lot. Thanks for good questions. I think this is the BEST questionnaire I have ever filled! Good luck with your work! Do not hesitate to contact me if you like to ask/discuss more. I would enjoy continuing in this kind of way, it makes me excited!"

"Human Language Project is an excellent initiative."

"Congratulations for the initiative and the option to include as many answers as possible."
"Best wishes to the survey - this is one of the most important topics for Europe at the present time."

4 Appendix I – Detailed analysis of open question answers

4.1 Questionnaire answer rates

	Questions	Provided an answer		Did not provide an answer	
		Absolute	Percentage	Absolute	Percentage
1	Personal details				
c	How many years of work experience do you have?	634	100%	0	0%
d	In which country are you based	631	99,53%	3	0,47%
2	What is the name of the organization you work for?	613	96,69%	21	3,31%
3	What type of organisation do you work for?	634	100%	0	0%
4	What is your company's estimated annual revenue in Euro?	69	10,9%	565	89,1%
5	What is the size of the organisation (total number of employees)?	631	99,53%	3	0,47%
6	What is your main role in the organisation?	634	100%	0	0%
7	What are the day-to-day responsibilities in your role?	623	98,26%	11	1,74%
8	What are the key research fields, areas and sub-areas, methods and applications you work on?				
a	Fields	625	98,58%	9	1,42%
b	Areas and sub-areas	618	97,48%	16	2,52%
c	Methods	596	94,01%	38	5,99%
d	Applications	616	97,16%	18	2,84%
9.1.	Which languages do you mainly work with in your research or offer in your products or services? A - M	598	94,32%	36	5,68%
9.2.	Which languages do you mainly work with in your research or offer in your products or services? N – Z	409	64,51%	225	35,49%
10.1.	Which languages would you like to include in your research, products or services in addition – but cannot due to a lack of technologies, tools or resources? A - M	302	47,63%	332	52,37%
10.1.	Which languages would you like to include in your research, products or services in addition – but cannot due to a lack of technologies, tools or resources? N – Z	279	44,01%	355	55,99%
11	In which of the following economic sectors do you see high potential for applications, opportunities for commercial growth or promising target markets for your research, products or services?	619	97,63%	15	2,37%

12	Which of the following Language Technology applications and services for the Multilingual Digital Single Market could be improved through your research?	605	95,43%	29	4,57%
13	Where do you see crucial gaps in terms of technologies, tools, or resources, especially with regard to specific languages?	407	64,20%	227	35,80%
14	What is the biggest challenge the European Language Technology community is currently facing?	408	64,35%	226	35,65%
	Visions for a future large-scale European LT Programme				
15	Do you support the idea of setting up a large-scale Human Language Project?	623	98,26%	11	1,74%
16	Are there any specific reasons why you do not support the setting up of a Human Language Project? Please specify if possible.				
17	The above-mentioned study suggests, in terms of the HLP's key strategic vision, to concentrate on achieving Deep Natural Language Understanding by 2030. Do you think this is the right vision and an adequate scientific challenge?	614	96,85%	20	3,15%
18	Which strategic vision would you suggest instead?				
19	How long do you think the HLP needs to be so that it can reach the suggested scientific vision and have a significant impact?	613	96,69%	21	3,31%
20	Let's assume that we have a comfortably funded Human Language Project with a timespan of ca. 5-15 years. What are, in your opinion, the (up to) five key challenges Europe needs to work in with regard to:				
a	Basic research	422	66,6%	212	33,4%
b	Applied research	343	54,10%	291	45,90%
c	Innovation	257	40,54%	377	59,46%
d	Industries/Sectors/Verticals	242	38,17%	392	61,83%
21	Which are the top three research, technology development, or socio-economic opportunities that you personally envisage the HLP to bring about or to successfully address?	337	53,15%	297	46,85%

22	Do you have any other additional suggestions or recommendations with regard to the HLP? For example, how it should be organised in terms of priority setting and governance or with regard to strategic guidance?	179	28,23%	455	71,77%
23	How should the Human Language Project be funded?	584	92,11%	50	7,89%
	What are, in your opinion, the five key topics, applications, services that must be included in such a platform?	301	47,48%	333	52,52%
25	Do you have any additional recommendations regarding the setup of the European Language Technology Data and Service Platform? For example, regarding the collaboration between data providers, LT providers and LT consumers?	143	22,56%	491	77,44%
	Talent generation and retention				
26	Which technical or soft skills do you personally consider most important for your specific area/projects?	454	71,61%	180	28,39%
27	How can the skill gap best be addressed?	560	88,33%	74	11,67%
28	Last but not least...				
a	How did you find out about this survey?	602	94,95%	32	5,05%
b	If you have any additional comments, concerns or suggestions please do not hesitate to share them.	525	82,808%	109	17,19%

4.2 Where do you see crucial gaps in terms of technologies, tools, or resources, especially with regard to specific languages?

Where do you see crucial gaps in terms of technologies, tools, or resources, especially with regard to specific languages?	Mentions
Insufficient research and availability of resources (especially for minority languages and dialects)	165
MT (especially for smaller languages)	42
Standard NLP systems are missing for most European languages	38
Insufficient funding is being made available, in particular for low resource languages	16
More domain specific resources are needed	15

Copyright restrictions (even for research use) pose challenges	14
Interoperability and standardisation need to be established	13
Semantics and pragmatics	12
Natural Language Understanding and processing	12
Text-to-speech and speech-to text applications	11
Educational tools	8
User models and applications with human-computer interaction	7
NLP applications and systems with more linguistic knowledge are needed	6
Multimodal input data is rarely processed	6
Integration of translation systems (translators still need to use too many stand-alone tools)	6
Collaboration between different research institutions	6
Dialog understanding	5
Applications using and producing simplified language	4
More cross-lingual work is needed	4
Multilingual Semantics: Aspect, Mood, Polarity, ...	4
Benchmarking availability	3
Industry cooperation	3
Tools for sign language users	2
GPUs	2
Conversational agents	1
Poor support of W3C / web standards for internet accessibility and interoperability	1
Not much research on voice output communication aids in Germany	1
Framework and API for unified access to syntactic representations for all languages	1
Proactivity for the development of applications focused on entrepreneurship and innovation in the Natural Language Processing area	1
Broad coverage language understanding	1
Language generation	1
Crowdsourcing is limited and not feasible for smaller languages	1

Many tools remain mostly static or they are released to the public and used as off-the-shelf	1
New assessment like done in the META-NET White Papers is needed	1
Post-GDPR personal data unification	1
Brain drain and immigration issues for non-EU expats pose challenges	1
Challenges are found in the cultural backgrounds and codifications of verbal and non-verbal communication	1
Telecommunications	1
Too little work being done on the relation of languages (intercultural communication)	1
Lack of large-scale computing facilities	1
Terminology	1
No proper method of binding ontologies exists	1
No solid theoretical basis for guiding the learning process of data-driven systems	1
Multilingual questionnaires for policy making , awareness of multilingualism in European politics are needed	1
More skilled individual need to improve resources for minority languages	1
Brexit	1
Argumentation Mining	1
Total	427

4.3 What is the biggest challenge the European Language Technology community is currently facing?

What is the biggest challenge the European Language Technology community is currently facing?	Mentions
Neglect of smaller languages and the importance of multilinguality	78
Lack of available data resources, especially for smaller languages	65
Unwillingness to collaborate (interdisciplinary cooperation), fragmentation of community no commonly defined goals	41
Lack of funding	41
MT quality	30

Dominance of English (there is little incentive to work on European languages other than English (e.g. in order to publish research on English needs to be included)	25
Competition with innovative systems and tools built in the US	22
Current limited awareness of LT potential for Europe	19
Not enough industry focus on development and use of applications	18
Lack of knowledge on Deep learning	13
Brexit	11
Lack of support from EC and member states	10
Quality of NLP tools	10
Insufficient results for semantic, pragmatic and discourse analysis	10
Licensing & Access	10
Standardization	9
Accessibility for diverse groups with special needs (e.g. elderly, disabled etc.)	8
Brain drain in academia	8
Understanding	7
Bureaucracy	4
Unbalanced economic development in countries	4
Multimodal processing	4
Better cross-lingual methods are needed	3
Too much focus on big data and machine learning, instead of the vastly superior rule-based approaches	3
More speech research needed	3
Multilingual adaptable interfaces are needed	3
Low profile of localisation/translation industry poses challenges	3
Lack of unified service layer which can support different components	2
Quality assurance metrics	2
User privacy concerns	2
Digital Single Market	2
Outdated software platforms	2

Interpretation of culture-specific linguistic metaphors	1
Balance between using real knowledge and possibilities of the virtual reality	1
Lack of true Venture Capital willing to take a chance	1
Novel cooperation between academia and industry	1
Common representation of linguistic phenomena for typologically different languages	1
Competition from China	1
Not having a representative language technology hub (as e.g. Montreal currently is for Deep Learning)	1
Ethics	1
Robustness	1
Focus should be shifted to economically rewarding languages like Chinese	1
Quality of IATE is falling behind	1
Importance of symbolic approaches	1
Cognitive modelling	1
Auto-tagging and machine learning accuracy	1
Keeping in line with actual customer needs	1
Generation of text	1
Computer power	1
High prices and bad quality of mobile Internet service providers	1
Integration	1
Mismatch between progress and knowledge in theoretical linguistics and its application in language technologies	1
Total	492

4.4 Are there any specific reasons why you do not support the setting up of a Human Language Project? Please specify if possible.

Reasons not to support a large-scale HLP	Mentions
Previous unsuccessful attempts of similar projects	4
Waste of money	3

Too much bureaucracy	2
Waste of resources	2
Might not benefit the whole community	2
Initiatives introduce bias into research strategies and methods	2
Lack of focus	1
Motivation	1
Unclear goals	1
These projects do not tend to lead through breakthroughs	1
Shared responsibility does not guarantee success (each country should take action)	1
Too early for such a large-scale project	1
Lack of already resources	1
Total	22

4.5 Which strategic vision would you suggest instead?

Which strategic vision would you suggest instead?	Mentions
Put focus on research, 2030 is too early for such a large-scale project	11
Priority should be on well-edited language resources	8
Language Technology poses challenges that cannot be solved with Understanding alone	7
Foster collaboration, education and development in smaller teams, support of individuals	6
More basic research needed (focus on NLP and giving transparency to black box)	5
Focus on Understanding and Generation	3
Strategic vision unnecessary	3
General setup of EU projects lack flexibility and agility needed for breakthroughs and innovation	3
Unique language model/clear definition needed before Understanding can be tackled	3
2030 too late, results are needed sooner in order to stay competitive	3
Research needed on new areas that capture more than just understanding	2

Bazaar instead of cathedral: Create a working marketplace of ideas with appropriate infrastructure (data, computing resources, funding) with smaller milestones	2
Reinventing Language Understanding through Crowdsourced Multi-level Cognitive and Functional Experimentation	1
Deep NLU tends to be used as a tool to supervise people. Our academy supports the use of an easy-to-learn, neutral second language for all.	1
Focus on enabling conversational interfaces (written or spoken) between humans and their environment	1
Shallow machine translation (that preserves content as supposed to SMT or NN)	1
Combine statistics and deep learning	1
More usable applications for end-users and integrating into larger IT-systems (information retrieval, different services for disabilities)	1
Acceptation of Natural Intelligence	1
Generate optimal assistance systems for speakers to operate adequately in a multilingual setting	1
Total	64

4.6 Let's assume that we have a comfortably funded Human Language Project with a timespan of ca. 5-15 years. What are, in your opinion, the (up to) five key challenges Europe needs to work in with regard to:

4.6.1 Basic research

Basic research	Mentions
Corpora/Resources/Ontologies/Dictionaries/Terminology repositories/Treebanks etc. with better annotations and effective legal framework	65
Deep learning/Neural networks/Brain research	47
NLP tasks and tools (e.g. including Q&A, WSD, Summarisation, Information extraction, Sentiment analysis etc.)	45
Natural Language Understanding	41
Cross-linguality/Multilinguality (for models, grammar formalisms and NLP applications such as QA and summarisation)	41
Machine Translation (including evaluation methods)	40
Focus on low-resource languages (this includes raising awareness for topic and assessing current state; from a technical point there should be more research on morphology, language-specific characteristics and NLP tools, deep learning	40

methods etc.)	
Semantics (incl. Meaning extraction	38
Sociolinguistics and Psycholinguistics (to assess what the social implications of a HLP would be, question of ethics, effects of social media, language acquisition etc.)	37
Linguistics (most frequently mentioned fields were Syntax, Morphology, Phonetics and Phonology, Formal grammars and challenges such as ambiguity in natural language	37
Knowledge and meaning modelling and integration (also domain-specific)	32
Speech (incl. ASR, structure of speech, speech production etc.)	23
Collaboration (refers to support from political players but also to interdisciplinary research and the definition of common goals and baselines)	21
Pragmatics	17
Data accessibility	16
Standardization and Integration of tools	15
Cross-modality	14
Reasoning	11
Funding	10
Neurolinguistics/Cognitive Modelling	10
Education (with availability of multilingual functionalities)	9
Generation	9
Artificial Intelligence	8
Evaluation metrics (scalable and objective, correlate with human judgement) with effective benchmarking and monitoring	7
Dialogue	7
Talent retention and generation	7
Inference	6
Domain adaptation	6
Sign language	5
Discourse	5
Rule-based methods	5
Big data	3

Prosody	3
Quantum computing	2
Data mining	2
Explanation	1
Machine Learning	1
Paralinguistic	1
Intelligence Amplification	1
Automata theory	1
Willingness	1
Transfer learning across languages	1
Conceptual graphs	1
Semantics	1
Detection of incongruences	1
Simultaneous parallel processing of natural languages	1
Fundamental principles of vocal interactivity between coupled agents based on ostensive inferential recursive mind-reading	1
ML (e.g., maxent methods based on stochastic relaxations)	1
Representation learning for large documents	1
Corpus based analysis	1
Robustness	1
User models	1
Weakly supervised learning	1
Dynamic human speech	1
Human machine interaction	1
Public services engagement (e.g. for translation memories)	1
Total	705

4.6.2 Applied research

Applied research	Mentions

MT (also including CAT tools)	64
Improve multilingual and interlinked resources/databases/terminology repositories/data representation (also for minority languages) and ensure standardization, interoperability as well as a legal environment for data sharing	48
Raise awareness for importance of LT and multilinguality (also on political level) to ensure funding	32
Improved open-source platforms with systems and applications as well as truly open and unencumbered data and code repositories	22
Conversational interfaces and agents	21
NLP and Linguistic baseline tools	20
Multimodal human machine interaction	19
Education and research on language acquisition	19
Domain adaptation	19
NLU	15
Dialogue systems	14
Collaboration between academia and industry to ensure community building. Strengthen Industry leadership on applied research and ensure better lab to market procedures	13
Speech technologies (incl. speech-to-text, text-to-speech, speech signalling, speech synthesis, cross language voice conversion, audio-visual biometrics, recognize speech from brain signals)	12
Integration of user feedback	11
ASR/Voice recognition	10
Summarisation	9
Meaning/Knowledge representation and extraction	9
Context aware systems (entailing pragmatic and semantic knowledge)	8
NLG	8
Question Answering	7
Semantics/Semantic web	7
Systems for inclusiveness (to support elderly, deaf, disabled etc.)	7
Fake news hate speech and fraud detection	6
Information extraction and retrieval	6
Improve evaluation processes and metrics	5

Personal assistants	5
Search Systems (Discovery systems esp. for research libraries)	5
Text and data mining	4
Social robotics	4
Deep learning	4
Massive language processing hardware architectures	3
Statistical interference	3
Argument mining; meeting, speech and legal document analysis	3
Discourse analysis	3
Sentiment analysis and opinion mining	3
More support for early stage researcher and students	3
Indexing	2
Seamless switching à la Google Home, Amazon Echo	2
Improve language resources/data representation	2
Unification of "basic" and "applied" research	2
Automatic search and annotation of images and annotation of language resources	2
Hybrid systems	2
AI	2
Analysis of large amounts of empirical data to test theoretical hypotheses	2
Software engineering	1
Applications in the area of language engineering	1
Syntax	1
Operational standards for LT	1
Field recording	1
Palantyping projects in every language	1
Development of friendly IT languages	1
Q&A	1
Hand-held devices. Interfaces via hand held devices (not just smartphones)	1
Universal Dependency treebank	1

Industry 4.0	1
Advanced voice-enabled automotive systems	1
Create a single European E-commerce environment without boundaries	1
Regulations	1
Virtual reality technology	1
Research on quantum computing	1
Web of Things	1
Robotics	1
Neuro-linguistic programming	1
Coordinated PoC's and a more long term support for the outcome of successful PoC's	1
Data formats unification	1
Computer Science	1
Interpretability of models	1
Machine learning	1
Total	491

4.6.3 Innovation

Innovation	Mentions
Inclusion of all languages and fostering of inter-cultural systems	28
MT (neural, spoken etc. with better pragmatic and semantic knowledge as well as predictive post-editing)	23
Better academia and industry relations and collaboration to seize business opportunities and foster start-up culture	22
Interactive NLP agents, personal trainers and assistants/dialogue systems	16
Bring together knowledge and methods developed for different fields and domains of NLT (e.g. e-government, e-health and e-justice and other public domains)	15
Make applications accessible to wider audience (including elderly, disabled etc.)	15
Visualizations and interfaces (also robust speech interfaces)	13
Language learner and educational systems and tools	13
NLU as well as SLU (innovate and build new tools)	12

Seamless human computer interaction	10
Open source framework (with packages and language models accessible through APIs)	9
Developing cognitive models of human-robots interaction	8
Neural networks, deep understanding and brain science	8
Voice/speech recognition and generation	8
Privacy issues, regulations and security	8
Multimodal Communication and NLP (incl. summarization, knowledge extraction etc.)	8
Funding and fostering of innovative research environments	6
Emotion-aware, thought and intention reading, creative thinking applications (incl. gaming, online support etc.)	6
Platform for research and experimental hypothesis testing/a grid computing infrastructure for European research institutions	5
Ensure usability and benefits for society	5
Fake news, hate speech detection and social media surveillance	5
Semantics	4
New evaluation measures and standards for research	4
More innovation (e.g. important are by-products that shorten the time-to-market of new innovative solutions (both products and services))	4
Better tools for language comprehension and generation (also multimodal)	4
Multimodal Communication	4
Meeting/discussion facilitation and analysis	4
Overcoming limitations of machine learning (unlearning/relearning/nuanced learning)	3
Visual Tagging (Automatic subtitling and annotation)	3
Semi-automated generation of resources and terminology	3
Linked data	3
More innovation in university curriculums	3
Virtual/Augmented reality technology	2
Free and public services/frameworks for translation and localization	2
Decision support systems	2

New conversational systems	2
More robust methods	2
Hybridization for rule-based models	2
Personalisation (for context-processing and recommendation)	2
Quantum learning and processing	2
Studies on effectiveness of novel approaches	1
(Self-driving) car related applications	1
Autonomous context aware cognitive systems	1
Collaborative robots for industry 4.0	1
Semiotic approaches regarding different modalities	1
Ways to enhance portable communication tools	1
Computational intention (intentional ELIZA)	1
Hardware demands (take climate change into consideration)	1
Analogue computing	1
Enhance business information management systems	1
Automatic Error Correction	1
Regular competitions and rewards for best ideas	1
Risk Aversion	1
Big data for NLP	1
Total	312

4.6.4 Industries/Sectors/Verticals

Industries/Sectors/Verticals	Mentions
Health	53
Education	33
Tourism and Travel	25
Law, Justice and Legal system	23
MT and other LT applications (also incl. post-editing and CAT tools)	21
E-Commerce	20

Entertainment, arts, creativity, culture and cultural heritage	20
Knowledge transfer and collaboration between academia and industry (e.g. UK's long running KTP (Knowledge Transfer Partnerships) or other joint research and PhD projects, e.g. CIFRE PHD in France)	20
Media	17
ALL industries, sectors and verticals should be included	17
Business and business services (incl. business intelligence applications)	15
Sharing of data and resources (also linked data) and ensuring data licensing	14
Multilingual devices and applications (especially for smaller languages)	13
Security (incl. Intelligence Defence, Cyber Security, Forensic and other military context)	12
Public services and administrations as well as social services	11
Government	11
Finance	8
Assistants, chatbots, conversational agents and social robotics	7
NLP	7
Communication and Telecommunication	7
Manufacturing and mechanical Engineering	7
Automotive industry	6
Citizens' participation (Especially including elderly and disabled)	6
Life sciences, biotechnology and biomedicine	5
Weaken dependence from US market and innovation	4
Interfaces	4
CRM	4
Insurance	4
Internet of Things	4
Publishing and journalism	3
Negotiation and argumentation mining	3
Robotics	3
Safety and crisis Management	3

IT components	3
Service infrastructure	3
Banking	2
Renewable energies	2
Agriculture	2
Voice recognition	2
Searching and annotating of images and videos	2
Privacy	2
Mobile applications	2
Gaming industry	2
Applications with semantic knowledge and analysis	2
Energy	2
Hospitality industry	2
HR Management	1
Deep learning	1
Deeper participation of military and intelligence communities. Learn from DARPA etc.	1
Data Science	1
Social Intelligence	1
Industry	1
Nanotechnologies	1
Environment and natural resources	1
Industry 4.0	1
Respect importance of smaller languages (increase government support)	1
Smart City	1
Ethics	1
Space	1
Health	1
Science	1

Multimodal communication and human-computer dialogue	1
Logistics	1
Making large existent industrial systems public	1
Total	456

4.7 Which are the top three research, technology development, or socio-economic opportunities that you personally envisage the HLP to bring about or to successfully address?

Which are the top 3 research, technology development, or socio-economic opportunities that you personally envisage the HLP to bring about or to successfully address?	Mentions
MT for all languages	76
Better access to multilingual data/service for all people (inclusiveness for minorities and people with special needs)	56
Multilinguality/Minority languages/remove barriers	47
Education and language learning	32
Health care	29
Improved communication/collaboration between stakeholders	25
Deep learning	24
More cultural awareness (as well as general awareness of the topic)	24
Understanding	23
Human machine interaction	22
Improved NLP tasks/systems	22
Conversational assistants/chat bots	20
New linguistic data/standardization	19
Semantics, pragmatics and discourse tasks	15
Advanced research in human brain processing/cognitive systems/neural computing	14
Cross-modal, cross-lingual analytics	14
Opinion mining/ sentiment analysis / social media analysis	11
Competitiveness with other global players	11

E-commerce	10
Interpersonal interfaces	10
Make LT services available for a wider community	9
Dialogue systems	8
Deployment of AI systems	8
NLP in robotics	7
Knowledge representation/modelling	7
Justice, Legal and Patent sector	7
Terminology	7
Speech synthesis	6
Summarisation	5
Speaker/speech recognition	5
Business	5
E-Government	5
Voice-based interaction	5
Public administration	5
Information extraction and retrieval	5
Innovation/ idea generation	5
Inference and reasoning	4
Machine learning	4
Language generation	3
Common platform	3
Reasoning and inference	3
World peace	3
Multilingual Digital Single Market	2
Establish ethical guidelines for HLT	2
Formal grammar	2
Fake news detection	2
Economic growth	2

More incentives for researchers to stay in Europe	2
Internet of Things	2
Funding	1
Social network analysis	1
Scalability	1
Business	1
Accelerated humanities research	1
Research Machine Learning/Statistical Methods	1
Cyber security	1
Finance	1
Management	1
Social semiotics	1
Human level ASR	1
Crowdsourcing	1
Real-time service provisioning	1
Energy efficiency of chips	1
Practical applications	1
Contradiction detection	1
Web of things	1
High quality services	1
Parsing	1
Integration of text geo-parsing technology	1
Total	657

4.8 Do you have any other additional suggestions or recommendations with regard to the HLP? For example, how it should be organised in terms of priority setting and governance or with regard to strategic guidance?

Do you have any other additional suggestions or recommendations with regard to the HLP? For example, how it should be organised in terms of priority setting and governance or with regard to strategic guidance?	Mentions
Include all European languages and foster protection of languages (in collaboration with member states and governments) and intercultural understanding. Ensure equality when it comes to resources etc.	52
Clear and standardized plan/agenda (guaranteeing inclusiveness and collaboration of research, industry, economics, politics) with pre-agreed milestones	25
Better funding policy, less bureaucratic, supporting also smaller-scale projects (also including project bidding)	17
Bottom-up approach starting with smaller goals	15
Favour more basic and applied research over innovation	12
Ensure democratic organisation through shifting president, committee and board members among institutions and countries	12
Repositioning of EU research to diversify from the US and large corporation paradigms. Focus on ideas and scientific breakthroughs	7
Foster collaboration of numerous stakeholders, allow for feedback (in form of surveys etc.)	7
Avoid the big operators and big consortia on EU level. Do not make international partners a prerequisite	6
Focus on and learn from user experience, pay attention to demands and incorporate customer service	6
Better school and university education, more focus on training and educational applications	5
Market driven approaches, ensuring business and industrial growth	5
Support early stage researchers and encourage the creation of local and regional research ecosystems	5
Include competitions, evaluation campaigns and other incentives (prize money) to foster sustainable development	4
Organisation driven by EU Commission	3
Development and assessment of deep learning and neural approaches	2
Focus on rule-based approaches	2

Communication through annual conference etc.	2
Consult and adopt best practices from similar projects (e.g. CITIA, FP6 Network of Excellence model)	2
Monitoring on ethics	2
Development of rich resources	1
Define a MVP service for one vertical	1
Organized by research institution in collaboration with industry	1
Individual language adaptation modules	1
Set up of preliminary board to conduct analysis and requirements	1
Collection of multilingual pathological speech databases	1
Adopt best practices from other projects (e.g. ERIC etc.)	1
Marketing and advertisement	1
Provide long-term co-funding of applied research departments that can do the idea-to-marked development.	1
Tools for machine translation	1
Set up industry peer groups	1
Goals are knowledge extraction and reasoning	1
Coaching in health systems	1
Open source initiatives and ease of access and integration	1
Include NGO's and the EFNIL (state languages) and NPLD (Network to Promote Linguistic Diversity)	1
Organisation by a consortium including non-specialists	1
Crowdsourcing language resources and analysis	1
Involve ELRA	1
Total	209

4.9 What are, in your opinion, the five key topics, applications, services that must be included in such a platform?

What are, in your opinion, the five key topics, applications, services that must be included in such a platform?	Mentions
MT (includes quality assurance, translation memory, CAT etc.)	121

Resources (Ontologies, lexicons, dictionaries etc.)	82
Education	49
NLP tools	48
Speech applications	44
Health	36
Easy accessibility	35
Corpus compilation	26
Information extraction and retrieval	23
Summarisation	22
E-participation and E-Government	22
Search systems	15
Intelligent Assistants	15
Semantic processing	14
Dashboards/usability and user feedback	14
Law and legal services	13
Dialogue systems and communication tools	13
Plug-in, download services and APIs	13
Multimodal applications	13
Understanding	12
Opinion mining and sentiment analysis	11
Q&A	10
Public Administration	9
Standards for interoperability	9
Domain adaptation	8
Cloud services	7
Transparency	7
E-commerce	7
NL generation	7
Fake news	6

Social media surveillance	6
Culture	6
NE detection and linking	6
Licensing	6
Machine learning	6
Training/MOOCS	5
Evaluation settings and benchmarks	5
Deep learning	5
Security	5
Text analytics	5
Data curation	5
Parsing	5
Interfaces	5
Open Algorithm Repository	4
Management and platform customer support	4
Business	3
AI	3
Sign language data	3
Marketing	3
Environment and agriculture	2
Finance	2
Platform should support open software projects like Gensim/NLTK/SpaCy	2
Competitions (along the lines of Kaggle)	2
Politics	2
Tourism	2
Automotive applications ("Alexa in the car")	2
E-Care	1
Computational Linguistics	1
Dispute resolution	1

Crowdsourcing platform	1
Education	1
Call centre services	1
Automated data format conversion	1
OCR	1
Industry	1
Privacy	1
Semantic processing	1
Energy	1
Human rights, feminism and gender bias	1
Transport	1
Multilingual templates for websites with elements of localization	1
Open Publication Infrastructure	1
Peace studies	1
NL search	1
News surveillance	1
Internet of Things	1
Travel and tourism	1
Research investment	1
Unix and Windows development environments	1
Robotics	1
Language preservation	1
Total	840

4.10 Do you have any other additional suggestions or recommendations with regard to the HLP? For example, how it should be organised in terms of priority setting and governance or with regard to strategic guidance?

Recommendations regarding the setup of the European LT Data and Service platform. For example, regarding the collaboration between data providers, LT providers and LT consumers.	Mentions
Easy accessibility of tools and data and open licensing	35
Involvement of all/various stakeholders	17
Unified, high-level, transparent and user-friendly approach with common goal	16
Data protection, exchange formats and standards	11
Ease collaboration between researchers in different countries (EU and outside) by providing better visa regulations etc.	8
Involve LT consumers	6
Educate and raise awareness for necessity of this project	4
Easier management and administration on EU level	3
Involve industry	3
Include knowledge structures, repositories of linked data and ontologies	3
Adopt best practices from CLARIN etc.	3
Involve data providers	2
Shift focus from academia, involve other stakeholders like industry	2
Enforce project evaluation processes	2
Involve private sector and translation sector	2
Creation of business models should be part of the project (e.g. Data Market Austria project: https://datamarket.at)	2
Inclusiveness and agility	2
Exploit already developed technology	2
Involve LT-Innovate, NPLD, British-Irish Council, Gala and TAUS	2
Commercialization plan, distribution and awareness is essential. Also share through new channels (news, TV, YouTube and other social media)	1
Link up with other European-funded projects	1

Release the parallel corpora from ELRC-SHARE for download without an account. Encourage bureaucracies to upload.	1
Encourage the machine learning aspect (avoid fundings for human-intensive work that will not be sustainable)	1
Discourage web services (except repositories or tooling related to workflow/application building - not execution)	1
Joint conferences and other dialogue platforms	1
Systematic datasets to be parsed in the hear, the think, and the speak mode	1
Creation of ecosystem to harvest use cases, and to demonstrate the value of adopting language based solutions	1
Domain adaptation	1
Define the target group, type of usage (human / machine) for each service vertical	1
Attention must be paid to the varying forms of "International Phonetic Alphabet"	1
Include minority languages	1
Develop the technology of human computer interaction so that it is grounded in the physical and social semantics of human discourse	1
Include speech	1
Layers based: text filtering, tokenisation, spelling, grammar and style checking, hyphenation, lemmatisation, parsing etc.	1
Inclusion of ISO TC37	1
Meta layered platform connecting everything that is already existing like the systemic Open Innovation PROMIS	1
Incorporate components for user feedback	1
Semantic Interoperability	1
Foster smaller, more focused projects	1
Support automatic intelligent analysis and schemes e.g. through nano-publications	1
Fund the organisation that runs this service sufficiently to provide the level of curation that is necessary for the platform to run well	1
The platform must be driven by DEMAND	1
Go to the facilities based carriers and get them as sponsors, both for computing capacity and specifically as real funding providers	1
Ability to plug-in proprietary or open application (along the lines of Apache LT tools, UIMA, GATE)	1

Hardware demands	1
Total	151

4.11 Which technical or soft skills do you personally consider most important for your specific area/projects?

Which technical or soft skills do you personally consider most important for your specific area/projects?	Mentions
Linguistics (including all disciplines: semantics, syntax, phonetics, formal linguistics, corpus linguistics etc.)	121
Programming (especially Python and Java, but also R, C++ and Perl)	109
Machine learning	84
Collaboration, team working and networking	62
Statistics	45
Math	45
Deep learning and neural networks	44
Innovative thinking, creativity and proactivity	43
Familiarity with NLP technologies	41
Language skills	37
Computer science	32
Willingness and eagerness to learn as well as curiosity	30
IT	24
Interdisciplinary work	22
Computational linguistics	20
Project management	19
Analytical thinking	18
Intercultural communication and knowledge	18
Scientific work style (writing and communication)	17
Adaptability, flexibility and ability to generalise	15
Technology enhanced learning and education	13

Logic and reasoning	13
Industry awareness to fill economic gaps	13
Software development	13
Data science	12
Software engineering	12
Localisation, translation and multilinguality	11
Speech processing	11
Hard working, perseverance and tenacity	10
Big data	9
User understanding and customer management	8
Python	8
Neuroscience and cognitive science	8
Empathy	7
CAT tools, ERP and post-editing	7
Data analysis	7
Artificial Intelligence	6
Attention to detail and meticulousness	6
Concepts machine understanding and human machine understanding	5
Compilation of data and resources	5
Knowledge representation and language resources	5
Negotiation and leadership skills	4
Shell scripting	4
Distributed systems	4
Terminology	3
Knowledge of data formats and encoding technologies (xml, RDF etc.)	2
Constructive criticism	2
Symbolic approaches	2
Courage to fail and critical awareness	2
GitHub, Docker, Forking	2

Coding	2
Passion for science	2
Language acquisition	2
Pattern recognition	2
Language and data modelling	2
C++	2
Computer vision for gesture and sign	2
Java	2
Open Licensing	1
Programming	1
Linguistics combined with graphic design	1
International trade and related law experience	1
Databases	1
Hardware infrastructure	1
AI	1
International telecommunications carrier and data transport experience	1
Services Architecture	1
System integration	1
Good work environment and perks (allowing remote work etc.)	1
Large scale computation	1
Social media experience	1
Web technologies	1
Social responsibility challenges (including privacy concerns and copyright infringement)	1
Eye tracking	1
Knowledge from sociology and psychology	1
Total	1093

5 Appendix II – Additional tables

5.1 In which country are you based?

In which country are you based?	Absolute	Percentage
Germany	75	12%
Spain	58	9%
UK	48	8%
France	44	7%
Italy	37	6%
Czech Republic	32	5%
Netherlands	25	4%
Belgium	20	3%
USA	18	3%
Greece	16	3%
Sweden	16	3%
Romania	15	2%
Hungary	14	2%
Ireland	14	2%
Lithuania	14	2%
Slovenia	14	2%
Denmark	13	2%
South Africa	13	2%
Poland	12	2%
Portugal	12	2%
Latvia	11	2%
Bulgaria	10	2%
Estonia	10	2%
Switzerland	10	2%
Austria	9	1%
Finland	9	1%
Serbia	8	1%
Iceland	7	1%
Russia	5	1%
Croatia	4	1%
Luxembourg	4	1%
Malta	3	0%
Norway	3	0%
Moldova	3	0%
Slovakia	3	0%
Canada	2	0%
China	2	0%
Israel	2	0%
Macedonia	2	0%

Thailand	2	0%	

Companies	Number of participants
Tilde	3
Nuance Communications	2
Palex	2
Oracle	2
Expert System	2
Phonexia	2
IAI Linguistic Content AG	2
inmark europa s.a.	2
Transmachina	1
Parisienne de photographie	1
Natural vox	1
Clementine	1
Syllabs	1
Company Tilde	
Vytautas Magnus University	1
Memsource	1
COMTEC Translations	1
BIK Terminology	1
Departmant of justice and corracttional services /freelance /	1
Clarify as	1
Deutsche Welle	1
TiP Sp. z o. o.	1
dhaxley Translations	1
Wolters Kluwer Deutschland GmbH	1
DialogCONNECTION Limited	1
Moravia	1
Everteam	1
Nlg	1
Appen	1
Bloomwise	1
Expert System	1
PROMIS@Service Sàrl	1
Freelance	1
Sunda Systems Oy	1
Geodan	1
text&form GmbH	1
Hermes Traducciones y Servicios Lingüísticos, SL	1
Top Communica	1
Homag	1
University College London	1
Humusha Translation Services	1
2Talk	1

0%		

Applied Logic Laboratory	1
Microsoft	1
IAI- Institute for Applied Information Science	1
Mozaika	1
IBM	1
Nico van de Water Linguistic Services	1
IBM	1
Artificial Solutions	1
IDIElkon	1
Orange	1
Rosoka Software	1
Pangeanic	1
SAIL LABS Technology	1
Carmeq GmbH	1
Seacastle	1
Prompsit Language Engineering	1
Sherpa	1
Softwin	1
SAP	1
Integris	1
Self	1
Intel	1
Siemens	1
Intel	1
Spaziodati	1
Interceptor Solutions Ltd	1
Supertext AG	1
IOLAR	1
SYSTRAN	1
ISVWorld	1
alinari	1
Textgain	1
K Dictionaries	1
TNT Express (FedEx)	1
Kaleidoscope GmbH	1
Translat, d.o.o.	1
KIE Srl	1
UAB TokenMill	1
Logical Events Limited	1
Web2Learn	1
Lucy Software Ibérica SL (a United Language Group company)	1
Wordbee	1
Mastervoice	1

Meltwater Group		
Total		
Angola		
Barbados	1	0%
Belarus	1	0%
Brazil	1	0%
Georgia	1	0%
India	1	0%
Iran	1	0%
Japan	1	0%
Nepal	1	0%
Pakistan	1	0%
Singapore	1	0%
Turkey	1	0%
Total	631	1

5.2 What is your company's estimated annual revenue in Euro?

Name of company	Annual revenue in Euros
IBM	77000000000
Microsoft	74448059376
TNT Express (FedEx)	50000000000
Oracle	37000000000
Intel	30000000000
Wolters Kluwer Deutschland GmbH	4300000000
Nuance Communications Deutschland GmbH	1950000000
Homag	1000000000
Deutsche Welle	302000000
Meltwater Group	200000000
IBM Italy	100000000
Moravia	100000000
Carmeq GmbH	60000000
inmark europa s.a.	40000000
Palex	3000000
Palex	3000000
Spaziodati	3000000

5.3 What is the size of the organisation (total number of employees)?

5.4 Which languages do you mainly work with in your research or offer in your products or services?

Which languages do you mainly work with in your research or offer in your products or services? Which languages do you mainly work with in your research or offer in your products or services?	Absolute	Percentage
English	536	18%
German	247	8%
French	224	7%
Spanish	201	7%
Italian	139	5%
Dutch	100	3%
Portuguese	98	3%
Russian	98	3%
Czech	81	3%
Other	78	3%
Arabic	76	2%
Polish	76	2%
Chinese	74	2%
Swedish	70	2%
Romanian	60	2%
Catalan	56	2%
Greek	56	2%
Hungarian	49	2%
Bulgarian	48	2%
Slovene	46	2%
Estonian	44	1%
Croatian	43	1%
Danish	43	1%
Finnish	42	1%
Lithuanian	42	1%
Slovak	42	1%
Turkish	42	1%
Latvian	37	1%
Norwegian (Bokmål)	37	1%
Serbian	32	1%
Hindi	28	1%
Icelandic	26	1%
Irish	26	1%
Norwegian (Nynorsk)	25	1%
Basque	24	1%
Urdu	21	1%
Galician	18	1%
Maltese	18	1%
Welsh	15	0%

Tamil	12	0%
Kurdish	9	0%
Scottish Gaelic	9	0%
Berber	3	0%
Total	3051	100%

5.5 Which languages would you like to include in your research, products or services in addition – but cannot due to a lack of technologies, tools or resources?

Which languages would you like to include in your research, products or services in addition – but cannot due to a lack of technologies, tools or resources?	Absolute	Percentage
German	81	5%
English	74	5%
Arabic	72	5%
French	71	5%
Chinese	68	4%
Spanish	64	4%
Other	63	4%
Russian	63	4%
Italian	56	4%
Czech	41	3%
Portuguese	41	3%
Hindi	38	2%
Turkish	38	2%
Polish	37	2%
Dutch	36	2%
Croatian	35	2%
Catalan	34	2%
Hungarian	34	2%
Swedish	33	2%
Serbian	29	2%
Danish	28	2%
Finnish	28	2%
Greek	28	2%
Lithuanian	28	2%
Slovene	28	2%
Welsh	28	2%
Romanian	27	2%
Maltese	25	2%
Bulgarian	24	2%

Tamil	24	2%
Urdu	24	2%
Basque	23	1%
Berber	23	1%
Estonian	23	1%
Irish	22	1%
Norwegian (Bokmål)	22	1%
Norwegian (Nynorsk)	22	1%
Slovak	22	1%
Icelandic	20	1%
Kurdish	20	1%
Galician	19	1%
Latvian	19	1%
Scottish Gaelic	19	1%
Total	1554	100%

5.6 What are the key research fields, areas and sub-areas, methods and applications you work on?

5.6.1 Fields

Fields	Absolute	Percentage
Language Technology	400	22%
Computational Linguistics	352	19%
Linguistics	262	14%
Artificial Intelligence	246	13%
Computer Science	191	10%
Data Science	153	8%
Cognitive Science	73	4%
Other	64	3%
Psycholinguistics	32	2%
Logic	24	1%
Mathematics	19	1%
Philosophy	15	1%
Neuroscience	7	0%
Total	1838	100%

5.6.2 Areas and sub-areas

Areas and sub-areas	Absolute	Percentage
Natural Language Processing	351	15%
Information Extraction	221	9%
Text Analytics and Text Mining	209	9%
Knowledge Representation and Knowledge Technologies	198	8%

Semantics	178	8%
Information Retrieval	174	7%
Machine Translation	167	7%
Language Technologies for Education	160	7%
Syntax	124	5%
Morphology	119	5%
Discourse	102	4%
Speech	94	4%
Multimodal Computing and Interaction	68	3%
Other	62	3%
Pragmatics	58	2%
Phonology and Phonetics	45	2%
Reasoning and Inference	36	2%
Total	2366	100%

5.6.3 Methods

Methods	Absolute	Percentage
Machine Learning	343	24%
Statistics	308	22%
Symbolic and Rule-Based Methods	291	20%
Deep Learning	236	17%
Cognitive, Mathematical, Computational Modelling	175	12%
Other	70	5%
Total	1423	100%

5.6.4 Applications

Applications	Absolute	Percentage
Multilingual Processing	247	9%
Document Analysis	243	9%
Resources, Corpus Development	225	8%
Machine Translation and Translation Aids	197	7%
Text Analytics and Text Mining	187	7%
Knowledge Management (incl. Semantic Web and Linked Data)	170	6%
Search and Information Retrieval	165	6%
Dialogue and Conversational Agents	129	5%
Interactive Systems	123	4%
Sentiment Analysis	121	4%
Grammatical Error Correction	115	4%
Text Categorization	105	4%
Natural Language Generation	102	4%
Social Media	100	4%
Speech Recognition, Text-to-Speech, Spoken Language Understanding	97	3%
Opinion Mining	96	3%
Question Answering	91	3%
Topic Modelling	59	2%
Speech Interfaces	58	2%
System Evaluation Methodology and Metrics	56	2%
Summarisation	51	2%
Paraphrasing	36	1%
Other	30	1%
Total	2803	100%

5.7 In which of the following economic sectors do you see high potential for applications, opportunities for commercial growth or promising target markets for your research, products or services?

In which of the following economic sectors do you see high potential for applications, opportunities for commercial growth or promising target markets for your research, products or services?	Absolute	Percentage
Education	439	15%
Information and Communication Technologies	399	13%
Human Health and Social Work	277	9%
Business Services	262	9%
Accommodation, Tourism and Food services	237	8%
Arts, Cultures and Other Services	234	8%
Justice and Legal	201	7%
Public Administration	194	6%

Entertainment	185	6%
Finance	144	5%
Industry and Manufacturing	117	4%
Insurance Industry	86	3%
Transportation, Logistics and Storage	85	3%
Trade and Repair	40	1%
Agriculture and Fishing	36	1%
Other	30	1%
Construction	22	1%
Total	2988	100%

5.8 Which of the following Language Technology applications and services for the Multilingual Digital Single Market could be improved through your research?

Which of the following Language Technology applications and services for the Multilingual Digital Single Market could be improved through your research?	Absolute	Percentage
Language Resources and Technologies (for Language Processing, Analysis and Production)	442	15%
Translation Services	279	10%
Multilingual Solutions for E-Learning	247	8%
Multilingual Solutions for E-Health	229	8%
Multilingual Content Curation and Production (including News and Media)	228	8%
Written and Spoken Language Interfaces (including, for example, Chatbots, Personal Assistants etc.)226	226	8%
E-Commerce (including, for example, Multilingual Product Catalogues)	222	8%
Social Intelligence (including, for example, Cross-lingual Opinion Mining and Sentiment Analysis on Social Media Data)	206	7%
Multilingual Solutions for E-Government	200	7%
Multilingual and Cross-lingual Customer Relationship Management including After Sales Management	179	6%
Multilingual Knowledge Graphs and Data Repositories	177	6%
Multilingual Solutions for E-Justice	158	5%
Cross-lingual Online Dispute Resolution	131	4%
Other	11	0%
Total	2935	100%

5.9 If you have any additional comments, concerns or suggestions please do not hesitate to share them.

If you have any additional comments, concerns or suggestions please do not hesitate to share them.	Mentions

Note of thanks	18
Good luck	18
Keep me updated (interest in getting involved)	13
Good work/nice survey (HLP is an excellent initiative etc.)	12
Survey too detailed	7
Survey took a lot of time	5
More funding for universities needed (project should not be driven by industry only)/give more incentives for work	5
Interdisciplinary and open-source approaches need to be encouraged, platform should be used by wide audience	4
EU doesn't handle language topic appropriately	3
Start bottom-up with a number of very small projects	3
More should be done for talent retention	3
Question were not relevant for my specific situation, to many "no opinion" or "not applicable" answers	2
The survey should be multilingual, definitions should be available	2
Copyright and patent laws should not restrict research, development and education (e.g. in health domain)	2
Research projects should not have funding as a primary goal	1
Test an already developed generic machine translation technology for English-Finnish (www.sunda.fi)	1
In order to survive we have to spend more resources on selling than developing.	1
Industry and academia should strengthen collaboration	1
Wales should be listed as a country in this survey	1
More research in neuroscience	1
Interesting tool in such a project might be an English style improving processor	1
Survey usability needs improvement	1
Necessary to formalize linguistic data with their meaning and content to interfere knowledge from the meaning	1
Too much focus on profit (increased focus of H2020 calls etc. on bettering infrastructure and industry is marginalising the smaller languages)	1
Member states need to be included to make any kind of joint effort more effective	1

Sceptical about reorganizing curricula initiative	1
Yoke the power of ICT to the expertise of linguists and pragmaticists	1
Support innovative start-ups in order to foster change	1
Look at ideas from the "Peace Machine" by AI researcher Timo Honkela	1
Project needs to be more clearly defined	1
Technology has to be in service of humanism	1
Interdisciplinary approaches need to be encouraged, platform should be used by wide audience	1
Reduce bureaucratic hurdles on EU level	1
Uncertain outlook on future collaboration because of Brexit	1
Regarding business probably product will be built by companies for the largest European languages, and the EU can help with resources to make this products available for smaller languages as well.	1
Worth considering setting up a one or two year part-time training programme for those in industry to upskill in the area of LT	1
Research programs should not only support main stream research	1
Project is restricted to theories, should be focus on concrete applications and results	1
Total	122

6 Presentation: Language Technologies for Multilingual Europe. Towards a Human Language Project

In the following, we include a presentation given by Georg Rehm at META-FORUM 2017 in Brussels, Belgium, on 13 November 2017.

Figure 4: Cover slide of the presentation "Language Technologies for Multilingual Europe"

META FORUM 2017

– Strategic Research and Innovation Agenda (V1.0) – Language Technologies for Multilingual Europe. Towards a Human Language Project

Georg Rehm

DFKI, Germany
georg.rehm@dfki.de

META-FORUM 2017
Brussels, Belgium – November 13/14, 2017

META-NET has received funding from the EU's Horizon 2020 research and innovation programme through the contract CRACKER (grant agreement no.: 645357). 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

META NET

- History – a brief look back
- SRIA Version 1.0 and the Human Language Project
- Survey "Language Technologies for Multilingual Europe"
- Conclusions

- **Published in early 2013.**
- **First strategic research agenda for our field.**
- **Complex process of collecting and shaping technology visions.**
- **Hundreds of researchers participated.**
- **Broad topics around Multilingual Europe in general.**

META NET

History

META-NET

- SRIA V0.5 presented at META-FORUM 2015 and Riga Summit.
- Built upon strategy papers and roadmaps prepared by several European projects, incl. the META-NET SRA (2013).

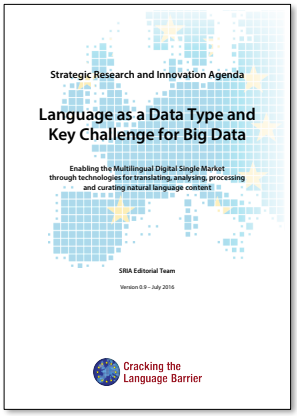
CRACKER
Cracking the Language Barrier for a Multilingual Single Digital Market

LT-OBSERVATORY

History

META-NET

- SRIA V0.9 unveiled at META-FORUM 2016
- Prepared, presented and endorsed by the Cracking the Language Barrier federation (editorial team).
- Explains how the LT community is going to make the DSM multilingual.



Cracking the Language Barrier

Application Areas in V0.9

META-NET

- **Multilingual E-commerce**
  - Customer-facing vs. back-office facing (after-market, after-sales)
  - Crosslingual search, CRM, helpdesks, processes, workflows
  - Semantic, crosslingual product descriptions and catalogues
  - Online dispute resolution
- **Multilingual Content, Media, Verticals**
  - Content analytics, curation, generation (incl. authoring support)
  - Multimodal communication (speech, written, IoT)
  - Vertical domains: health, government, mobility, energy, legal.
- **Translation, Language, Knowledge, Data**
  - Translation Centre – written/spoken, automatic/human
  - Crosslingual public and social intelligence, business intelligence
  - HQ resources, under-resourced languages, domain-specific LRs

- **STOA Workshop in European Parliament** (January 2017): "Language equality in the digital age – towards a Human Language Project"
- **Human Language Project** vision suggested in several presentations
- STOA Study, published in March 2017, recommend setting up the HLP

➢ http://www.stoa.europarl.europa.eu/stoa/cms/home/workshops/language

Current Developments

- **Multilingual Europe:** our languages enjoy equal status yet digital extinction of the majority of EU languages is a very severe danger.
- **Language Technology Research and Innovation in Europe:** World class research results (e.g., in QT21), strong SME base, thousands of LSPs; fragmentation; need for coordination.
- **Digitisation of our Continent – Big need for HQ Language Technologies:** translation, personal assistants, MDSM etc.
- **Artificial Intelligence:** Important breakthroughs and massive investments in R&D and applications (mostly in the US and Asia) – huge opportunity for Europe!
- **The European Language Challenge** cannot be abandoned or outsourced!

➢ **Need for Language Technology made *in* Europe *for* Europe!**

SRIA Version 1.0 beta

- SRIA V1.0 beta – unveiled today at META-FORUM 2017
- Prepared and presented by Cracking the Language Barrier federation
- Extended editorial team
- Document available on
  http://www.cracker-project.eu
  http://www.cracking-the-language-barrier.eu

Cracking the Language Barrier

SRIA Version 1.0 beta

META NET

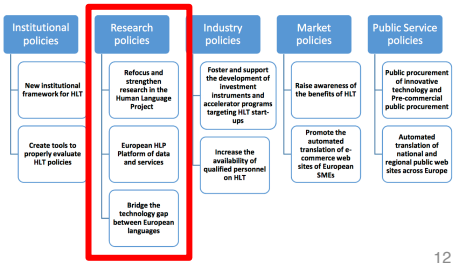
- SRIA 1.0 beta fully endorses and supports the STOA study
- Also our key recommendation: set up the **Human Language Project**
- Positive side effect: establish the Multilingual Digital Single Market
- SRIA informed by survey "Language Technology for Multilingual Europe"

Cracking the Language Barrier

STOA Study

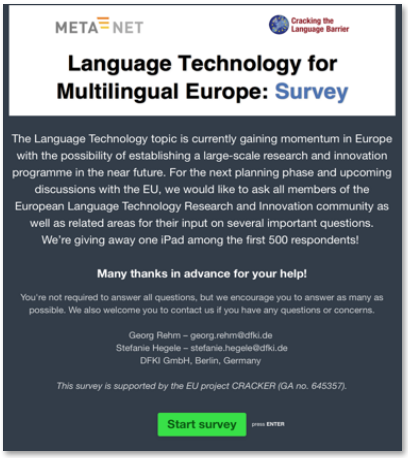
META NET

- "Language Equality in the Digital Age – Towards a Human Language Project"
- Strongly recommends to initiate the HLP
- Five main policy groups, 11 recommendations
- Research policies
  - Refocus and strengthen research in the HLP
  - European HLP Platform of data and services
  - Bridge the technology gap between European languages



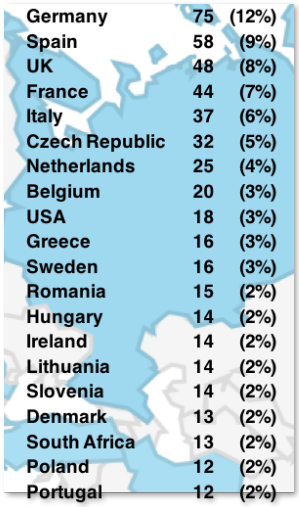
Multilingual Europe Survey

- Conducted in May/June 2017
- 29 questions (16 open, 13 m/c)
- Three main parts:
  1) Background and interests
  2) Visions for a large-scale HLP
  3) Talent generation/retention
- 634 participants from 52 countries
- *Very* high completion rate (27%)
- Avg. time to complete: 35,48 mins.
- ➔ Respondents are extremely passionate about this topic!



Survey: Demographics

- 634 respondents from 52 countries (37 European countries, 27 MS)
- High level of seniority (53% have >20 years of experience, 27% >10 years)
- Diverse Expertise: Language Technology, Computational Linguistics, Linguistics, Artificial Intelligence, Computer Science
- Majority based at universities (63%)
- Substantial group of participants (16%) from industry



Survey Results 1/7

- **Support for a Human Language Project**

97% of all respondents support the idea of a

HUMAN LANGUAGE PROJECT

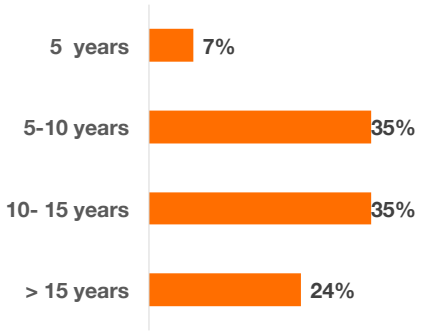
87% of all respondents believe

DEEP NATURAL LANGUAGE UNDERSTANDING AND GENERATION BY 2030

to be an adequate scientific challenge and goal

Survey Results 2/7

- **Timeframe and Stakeholder Involvement (funding)**



Stakeholder	Votes
European Commission	89%
Member States	57%
Industry	57%

Shared Programme between the EU and the Member States!

Survey Results 3/7

- **Economic Sectors – Impact of Language Technologies, especially in the context of the DSM**

Sectors to have the highest potential contribution to commercial growth	
Education	71%
Information and Communication Technologies	64%
Human Health and Social Work	45%

Specific services and applications that could benefit the Multilingual DSM	
Language Resources and Technologies	73%
Translation Services	46%
Multilingual Solutions for E-Learning	41%
E-Health	38%

Survey Results 4/7

- **Key Research Areas mentioned by the respondents**
  - Deep Learning, Natural Language Understanding
  - Data and Knowledge Repositories
  - NLP Applications
  - Machine Translation
  - Speech
  - Open Source Platforms
  - Conversational Interfaces and Agents
  - Multimodal Human Machine Interaction

Survey Results 5/7

- **Applications**

  Machine Translation

  Download services for multilingual resources including ontologies, lexicons, dictionaries etc.

  More in-depth development of NLP tools is encouraged, especially speech applications

  Other applications include IE and IR, summarisation, search systems and intelligent assistants

Survey Results 6/7

- **European LT Data and Service Platform**

  Importance of easy accessibility and open licensing for available tools and data including commonly agreed upon exchange formats and standards (30%)

  Involvement of all stakeholders, i.e., data providers, LT providers and LT consumers (11%)

  Unified, high-level, transparent and user-friendly approach with common goals (11%)

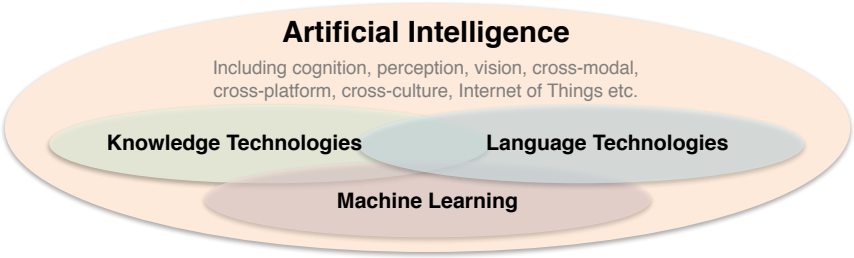
Survey Results 7/7

- **Talent Generation and Retention**

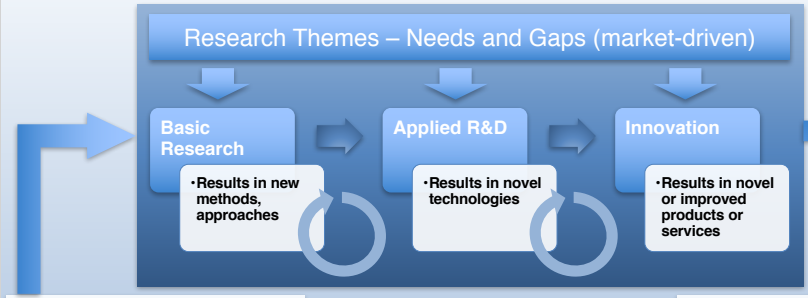
Closer collaboration between academia and industry (e.g., through job fairs, hackathons etc.)	74%
Reorganisation of university curriculums	62%
Fostering a more enterpreneurial culture (e.g., through specialised course modules, accelerator programmes etc.)	43%

Human Language Project

- **Large-scale EU funding programme – 10-15 years**
- **Goal: Deep Natural Language Understanding by 2030**
- **Artificial Intelligence for Next Generation Language Technology!**
- **New breakthroughs and groundbreaking results for industry, society, innovation, economy (multilingual digital single market).**



Human Language Project – Interdisciplinary R&D&I Programme



- Computational Linguistics
- Artificial Intelligence
- Language Technology
- Linguistics
- Computer Science
- Cognitive Science
- *other related fields*

HLP: Umbrella programme to turbo-charge and to coordinate all European R&D&I activities in a systematic way including EP, EC, Member States.

- New, groundbreaking methods, paradigms, approaches
- Foster technologies, products, innovation, economy
- Foster education

23

Human Language Project

META-NET

- **Goal:** Deep Natural Language Understanding
- **Breakthroughs in Artificial Intelligence plus a fresh look at Linguistics for the Next Generation of LT!**
- **All official European and many additional languages**
- **Broad coverage, high quality, high precision**
- **Create approaches, algorithms, data sets, resources**
- **Across modalities:** text, text types, speech, image, video etc.
- **Across platforms:** messaging, telephony, social, mobile, IoT etc.
- **Across cultures:** knowledge, customs, formalities, humour, emotion, subjectivity, biases, opinions, filter bubble etc.

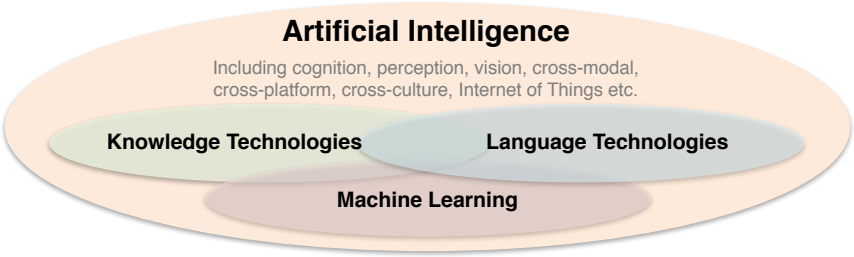
Human Language Project

META NET

- **Collaboration and coordination** between EC, EP, Member States and all other stakeholders.
- **Mix of funding sources:**
  - EU projects: Horizon 2020 (WP 2018-2020) + FP9 (2021+)
  - National/regional funding sources
- **Setup:** basic research, applied research, innovation, commercialisation – tightly intertwined
- **Timeframe:** at least 10 years
- **Policy change** towards "LT-enabled multilingualism"
- **Public procurement:** EU/EC, MS administrations should demand certain language technologies

Key Ingredients

META NET



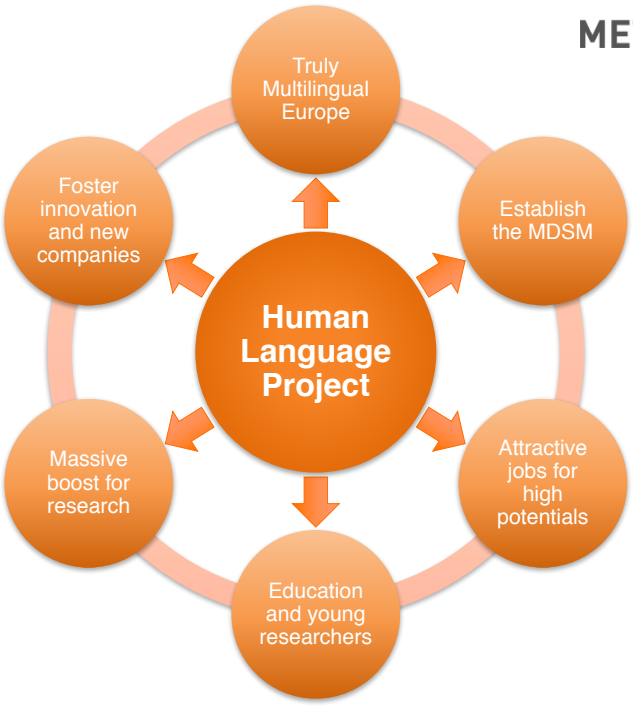
- Extend knowledge bases
- Semantic Web, ontologies, linked data, interoperability
- More complex models
- Multilingual resources that are grounded, extensible
- Subjectivity, objectivity, further novel dimensions
- Web-scale reasoning

- Combine DNNs and symbolic processing
- ML for knowledge acquisition and extension
- DNNs embedded into modular systems including symbolic knowledge bases
- Make it possible to inspect and also to optimise DNNs (beyond end-to-end)

- (Computational) Linguistics research towards deep language understanding
- From corpora to DNNs to annotated data to highly improved symbolic methods
- Language portability
- *Full and Deep Language Understanding by 2030 –* ***Human Language Project***

HLP: Selected Topics

- **High-Quality MT** – overcome quality (and language) barriers, written and spoken, collaborate closely with human translators
- **Content Curation** and Smart Online Content
  - Increasing commercial and social relevance of content ("fake news")
  - Include: domain, text type, style, register, discourse, social etc.
  - Type-specific, genre-specific analysis, assessment, generation
- **Multilingual European Knowledge Graph** that consolidates existing and emerging data (for crosslingual search, BI etc.)
- **Conversational interfaces**, especially for IoT, WoT, Industrie 4.0
- **Multilingual Europe**: LRs and LTs for *all* European languages
  - Include Member States – make it *coordinated*, *shared*, *focused*
  - Set of basic tools as open source and SaaS (free of charge)
  - Goal: boost the LT ecosystem and MDSM



Conclusions

- The Human Language Project is an opportunity for Europe to invest in a promising and sustainable field
- A topic that Europe can shape and claim as its own
- Investing into the HLP would solve the threat of Digital Language Extinction
- Investing into the HLP would make the DSM multilingual
- Investing into the HLP would secure Europe's place in the pole position in this field for many years to come

Next Steps

- The current SRIA is version 1.0 beta
- At META-FORUM 2017 we collect a final round of feedback and input
- We will consolidate all feedback and input afterwards and publish the final SRIA 1.0 in December 2017
- Important and relevant: the situation and discussion in the EP regarding the planned own-ini study
- Crucial goal for 2018: influence the discussion of the key pillars of FP9

META FORUM 2017

Thank you!

office@meta-net.eu

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