



Language equality in the digital age

Towards a Human Language Project

STUDY

Science and Technology Options Assessment

EPRS | European Parliamentary Research Service

Scientific Foresight Unit (STOA)

PE 581.621

Language equality in the digital age - Towards a Human Language Project

Study

IP/G/STOA/FWC/2013-001/Lot4/C2

March 2017

Abstract

In the digital era, language barriers represent a major challenge preventing European citizens and businesses from fully benefiting from a truly integrated Europe. These barriers particularly affect the less educated and older population, as well as speakers of smaller and minority languages, thus creating a notable language divide. Language barriers have a profound effect on (1) cross-border public services, (2) fostering a common European identity, (3) workers' mobility, and (4) cross-border e-commerce and trade, in the context of a Digital Single Market.

The emergence of new technological approaches such as deep-learning neural networks, based on increased computational power and access to sizeable amounts of data, are making Human Language Technologies (HLT) a real solution to overcoming language barriers. However, several factors, such as market fragmentation, uncoordinated research and insufficient funding, are hindering the European HLT industry, while putting under-resourced languages in danger of digital extinction.

Moreover, language technologies are not properly represented in the agenda of European policy-makers, although they are likely to be crucial for the construction of a fair and truly integrated European Union.

Based on the analysis of the current state of affairs, we argue for setting up a multidisciplinary large-scale coordinated initiative, the European Human Language Project (HLP). Within the HLP, eleven policies are proposed and assessed. These policies are grouped into: institutional policies, research policies, industry policies, market policies, and public service policies.

The STOA project 'Language equality in the digital age: towards a Human Language Project' was carried out by Iclaves with the support of the Universitat Pompeu Fabra at the request of the Science and Technology Options Assessment Panel, and managed by the Scientific Foresight Unit (STOA) within the Directorate-General for Parliamentary Research Services (DG EPRS) of the European Parliament.

AUTHORS

Rafael RIVERA PASTOR, Iclaves S.L.

Carlota TARÍN QUIRÓS, Iclaves S.L.

Juan Pablo VILLAR GARCÍA, Iclaves S.L.

Prof. Toni BADIA CARDÚS, PhD, Universitat Pompeu Fabra

Prof. Maite MELERO NOGUÉS, PhD, Universitat Pompeu Fabra

Acknowledgments

Several experts and organizations reviewed the study or were interviewed during the project and the authors would like to thank them for their valuable ideas and contributions to this report. The list of experts and organizations is included in Annex 7.1. The authors would also like to thank the R Development Core Team for providing the tool used in the quantitative analysis (R Core Team, 2016b), the developers of the R-studio environment (Studio, 2012) and all the developers of the R libraries used in the analysis and cited in Annex 7.1.

STOA RESPONSIBLE ADMINISTRATOR

Zsolt G. PATAKI

Scientific Foresight Unit (STOA)

Directorate for Impact Assessment and European Added Value

Directorate-General for Parliamentary Research Services

European Parliament, Rue Wiertz 60, B-1047 Brussels

E-mail: zsolt.pataki@europarl.europa.eu

LINGUISTIC VERSION

Original: EN

ABOUT THE PUBLISHER

To contact STOA or to subscribe to its newsletter please write to: STOA@ep.europa.eu

This document is available on the Internet at: <http://www.ep.europa.eu/stoa/>

Manuscript completed in March 2017

Brussels, © European Union, 2017

DISCLAIMER

This document is prepared for, and addressed to, the Members and staff of the European Parliament as background material to assist them in their parliamentary work. The content of the document is the sole responsibility of its author(s) and any opinions expressed herein should not be taken to represent an official position of the Parliament.

Reproduction and translation for non-commercial purposes are authorised, provided the source is acknowledged and the European Parliament is given prior notice and sent a copy.

PE 598.621

ISBN 978-92-846-0698-6

doi : 10.2861/136527

QA-02-17-247-EN-N

Table of contents

List of abbreviations.....	5
List of tables	6
List of figures.....	8
Executive summary.....	11
1 Introduction.....	17
2 Methodology and resources used.....	18
3 Synthesis of the research work and findings.....	19
3.1 Human Language Technologies in multilingual europe.....	19
3.1.1 Human Language Technologies: an overview.....	19
3.1.2 Linguistic resources and data.....	26
3.1.3 Technology gap between English and the other languages.....	29
3.1.4 LT-related research and industry in Europe	38
3.2 Socio-economic implications of multilingualism.....	45
3.2.1 Socio-demographic consequences of a non-multilingual regime in a multilingual Europe	46
3.2.2 Consequences of a non-multilingual regime within the countries for migrants and cross-border mobility.....	51
3.2.3 The impact of non-multilingual Europe on providing public services.....	53
3.2.4 A truly multilingual Europe fostering the European construction, citizenship engagement and reinforcing a common identity.....	58
3.2.5 The effect of a non-multilingual Digital Single Market in cross-border e-commerce.....	59
3.2.6 Impact on SMEs.....	66
3.2.7 Other effects of multilingualism on cross-border trade and businesses in Europe	68
3.3 Human Language Technologies and public policies	68
3.3.1 Trends in multilingualism and HLT in the EU bodies.....	70
3.3.2 EU multilingualism and HLT policies.....	73
3.3.3 Multilingualism in national and regional policies.....	77
3.3.4 Multilingualism policies outside EU	78
3.3.5 Recommendations of the industry and research in the LT community	79
4 Policy options.....	87
4.1 Challenges	89
4.1.1 Institutional challenges.....	89
4.1.2 Social challenges.....	89

4.1.3	Economic challenges.....	90
4.1.4	Sector challenges	91
4.2	Assessment criteria	91
4.3	Institutional policies.....	92
4.3.1	Reinforce the role of HLT within the institutional framework of multilingualism related bodies	92
4.3.2	Create tools to properly evaluate HLT policies	94
4.4	Research policies	96
4.4.1	Refocus and strengthen research in LT through a Human Language Project	96
4.4.2	Promote the European LT Platform of data and services.....	99
4.4.3	Bridge the technology gap between European languages.....	102
4.5	Industry policies	104
4.5.1	Foster and support the development of investment instruments and accelerator programs targeting HLT start-ups.....	104
4.5.2	Increase the availability of qualified personnel on HLT.....	106
4.6	Market policies	107
4.6.1	Raise awareness of the benefits for companies, public bodies, and citizens of the availability of on-line services, contents and products in multiple languages.....	107
4.6.2	Promote the automated translation of e-commerce websites of European SMEs.....	109
4.7	Public service policies.....	111
4.7.1	Public procurement of innovative technology and pre-commercial public procurement	111
4.7.2	Foster the translation of national and regional public web-sites and documents to other EU languages by using HLT	113
5	Conclusions	115
6	References	119
7	Annexes	129

List of abbreviations

AI	Artificial Intelligence
ASR	Automatic Speech Recognition
DAE	Digital Agenda for Europe
DSM	Digital Single Market
EC	European Commission
EP	European Parliament
EU	European Union
GDP	Gross Domestic Product
HLP	Human Language Project
HLT	Human Language Technologies
ICT	Information and Communication Technologies
IP	Internet Protocol
ISP	Internet Service Provider
LT	Language Technologies
MT	Machine Translation
NLP	Natural Language Processing
NMT	Neural Machine Translation
SM	Single Market
SMT	Statistical Machine Translation
TTS	Text-to-Speech Technologies

List of tables

Table 1: Common NLP tasks	21
Table 2: Languages included in the updated cross-language comparison	31
Table 3: State of LT support for 30 European languages in four different areas	33
Table 4: Leading technology suppliers in the global market	42
Table 5: LT challenges and required actions	81
Table 6: Assessment matrix for the "Reinforce the role of HLT within the institutional framework of multilingualism related bodies" policy	93
Table 7: Assessment matrix for the "Create tools to properly evaluate HLT policies" policy	95
Table 8: Assessment matrix for the "Refocus and strengthen research in HLT" policy	98
Table 9: Assessment matrix for the "Promote the European LT Platform of data and services" policy	101
Table 10: Assessment matrix for the "Bridge the technology gap between European languages" policy	103
Table 11: Assessment matrix for the "Foster and support the development of investment instruments and accelerator programs targeting HLT start-ups" policy	105
Table 12: Assessment matrix for the "Increase the availability of qualified personnel on HLT" policy	106
Table 13: Assessment matrix for the "Raise awareness of the benefits for companies, public bodies, and citizens of the availability of on-line services, contents and products in multiple languages" policy	108
Table 14: Assessment matrix for the "Promote the automated translation of e-commerce web-sites of European SMEs" policy	110
Table 15: Assessment matrix for the "Public procurement of innovative technology and Pre-commercial public procurement " policy	112
Table 16: Assessment matrix for the "Foster the translation of national and regional public webs and documents to other EU languages by using HLT" policy	114
Table 17: Summary of the main socio-economic consequences of language barriers analysed in the study	115
Table 18: Summary of the policy options of the HLP	118
Table 19: Official languages by country	131
Table 20: LOGIT Regression results	133
Table 21: Regression results for cross-border workers' mobility using Robust Standard Errors	142
Table 22: Regression results for cross-border workers' mobility using Bootstrapping (10 000 iterations)	143
Table 23: Regression results for cross-border on-line shoppers	146
Table 24: Regression results for cross-border on-line shoppers using Bootstrapping (10 000 iterations)	147
Table 25: Regression results for cross-border on-line retailers	151
Table 26: Regression results for cross-border on-line retailers using Bootstrapping (10 000 iterations)	152

Table 27: HLT related projects FP7	161
Table 28: HLT related projects H2020	162
Table 29: HLT related projects ICT-PSP	162

List of figures

Figure 1: Language Technologies: tools and resources.....	20
Figure 2: Untagged version of a fragment of the Brown corpus compared to a version annotated with part-of-speech tags	27
Figure 3: Annotated sentence of the Penn TreeBank.....	27
Figure 4: Screen shot of an ELAN file.....	28
Figure 5: Level of support of MT by language.....	34
Figure 6: Language usage in websites (percentage of websites)	35
Figure 7: Languages supported in selected Internet services	37
Figure 8: Worldwide LT market forecast (billion euro)	41
Figure 9: LT revenues by regions (percentage)	41
Figure 10: Location by EU region of LT industry, globally and by sectors.....	43
Figure 11: LT European industry by size (number of employees)	43
Figure 12: European LT vendors by segment.....	44
Figure 13: Percentage of population not speaking English by country	47
Figure 14: Percentage of population not speaking English by different socio-demographic factors	48
Figure 15: Evolution of percentage population in 5 year cohorts not speaking English by level	49
Figure 16: Percentage of population not speaking English, German, French, Italian, Polish or Spanish (at least good) by country	49
Figure 17: Percentage of population not speaking English, German, French, Italian, Polish or Spanish by different socio-demographic factors	50
Figure 18: Percentage of population not speaking any of the official language(s) of the country by country	52
Figure 19: Percentage of population that have lived and worked in another EU country (current situation compared to having no language barriers)	53
Figure 20: Disenfranchisement rate when introducing a new language (logarithm scale)	56
Figure 21: Yearly cost per disenfranchised person per language (logarithm scale)	57
Figure 22: Cross-border e-shoppers evolution.....	59
Figure 23: Cross-border e-sellers evolution.....	60
Figure 24: Cross-border language barriers for consumers by country	60
Figure 25: Relationship between language barrier and percentage of buyers.....	61
Figure 26: E-commerce traffic broken down by source	61
Figure 27: Cross-border e-commerce barriers for enterprises selling on-line to individual consumers not selling to EU but trying or considering (percentage of companies)	62
Figure 28: Clusters of cross-border e-commerce (consumers' side)	63
Figure 29: Descriptive regression results for cross-border on-line shoppers.....	64

Figure 30: Population shopping on-line in Hungary from other countries depending on language barriers	65
Figure 31: Cross-border e-commerce language barrier for enterprises selling on-line to individual consumers that are trying or considering to sell abroad by size	67
Figure 32: Companies doing electronic sales to other EU countries by size	67
Figure 33: Evolution of number of pages translated by source language (percentage of pages; total in millions)	70
Figure 34: Average percentage of documents including the topics (language technology topics compared to other trending technology topics)	72
Figure 35: Average number of sentences per document including the topic (language technology topics compared to other trending technology topics)	73
Figure 36: Official EU documents and their relevance to the topic based on number of sentences by type of document and year	74
Figure 37: Non-technological challenges of the LT Industry	80
Figure 38: Impact of LT on horizontal segments	82
Figure 39: The Multilingual Value Programme	84
Figure 40: Diagram of the European Platform for the Multilingual Digital Single Market	86
Figure 41: The European Human Language Project	87
Figure 42: The European Human Language Project and proposed policies	88
Figure 43: Proposed HLT policies	89
Figure 44: HLP Research strategy	97
Figure 45: Dendrogram of clusters of cross-border e-commerce (consumers' side)	134
Figure 46: Average percentage of population of the origin country speaking the language of the destination country by cluster (consumers' side)	134
Figure 47: Comparison of within and between average cross-border e-commerce by cluster (consumers' side)	135
Figure 48: Cross-border e-commerce by cluster (consumers' side)	136
Figure 49: Dendrogram of clusters of cross-border e-commerce (retailers' side)	137
Figure 50: Clusters of cross-border e-commerce (retailers' side)	137
Figure 51: Comparison of within and between average cross-border e-commerce by cluster (retailers' side)	138
Figure 52: Cross-border e-commerce by cluster (retailers' side)	139
Figure 53: Percentage of cross-border online-shoppers of total on-line shoppers by cluster and depending on language barriers (box-plot and mean)	148
Figure 54: Country average of the percentage of cross-border online-shops of total on-line shops by cluster and depending on language barriers (box-plot and mean)	153
Figure 55: Number of reports of the DG CONNECT by term and year	156
Figure 56: Number of blogs of the Digital Single Market of the DG Connect by concept and year	157
Figure 57: Number of blogs of the European Parliamentary Research Service by concept and year	158

Figure 58: List of Official EU documents analysed and number of sentences including the topic.....	159
Figure 59: Annual contribution of HLT related projects FP7	163
Figure 60: Per capita EU fund contribution by country for Competitiveness of SMEs (depending on language barriers with majority languages).....	164
Figure 61: Per capita EU fund contribution by country for Efficient Public Administration (depending on language barriers with majority languages)	165

Executive summary

In the European Union (EU) there are 24 official languages and more than 60 national and regional languages. Multilingualism represents one of the greatest assets of cultural diversity in Europe and, at the same time, one of the most substantial challenges for the creation of a truly integrated EU. Overcoming language barriers becomes crucial for the EU in the digital era. European citizens need to communicate in their own languages across the borders of Europe in order to increase workers' mobility, access to European public and private services and contents, and seize the opportunities of the Digital Single Market (DSM). This is a formidable challenge, which Human Language Technologies (HLT) can undoubtedly contribute to addressing.

HLT are key to overcome language barriers

Human Language Technologies are found behind many everyday digital products, since most of them use language to some extent. Mobile communications, social media, intelligent assistants, and speech-based interfaces are transforming the way citizens, companies and public administrations interact in the digital world. HLT are critical pieces of technology to aid in this digital revolution.

The emergence of new approaches such as deep-learning neural networks, based on increased computational power and access to huge amounts of data, are making HLT a real solution to overcome language barriers. The increasing availability of high quality translation and real-time speech recognition services are creating the tools for businesses and public bodies to seize the opportunity to provide their services, contents, and support in any language.

Improvements in HLT rely mainly on the ability to access and maintain ever larger and more finely-tuned linguistic data and resources. Lack of access to data is constraining the technological development of HLT. Acquiring and using it relies on cooperation between the industry and the different constituencies that generate, own, need and use this data. Thus, close collaboration between the industry and data owners becomes a necessity. Moreover, regulation of the use of such data should be made much more open and core language resources (annotated corpora, lexicons and ontologies) should be made interoperable and shared in an open environment.

Specific challenges in Europe for minority (or less-resourced) languages and the HLT sector

According to the European Charter for Regional or Minority Languages, *regional or minority languages* are languages that differ from the official language(s) of the state. These languages are traditionally used within a given territory of a state by nationals who form a group numerically smaller than the rest of the state's population. Therefore, by definition, they are in opposition to *official languages*. Although it is true that, in general, minority languages are at a disadvantage in terms of diffusion and support compared to official languages, it is preferable to refer to them as *smaller languages* (i.e. languages with fewer speakers) or, more directly, *under-resourced languages*.

Although smaller or minority languages are the ones to gain most from language technologies, tools and resources for them tend to be scarce – in some cases almost non-existent. In fact, there is a widening technology gap between English and the other official, co-official or non-official EU languages, some of which might already be facing digital extinction. In order to bridge this technology gap, policies should focus on fostering technology development for European languages other than English, particularly the smaller and less-resourced ones, therefore promoting language preservation through digital means.

Europe has a strong scientific base in language engineering and technology, and there is no shortage of innovative new entrants. While European R&D has produced a steady stream of small LT-based companies, fragmented European industry is not able to effectively respond to current technology

challenges. There are some top players in language technologies, such as Google and Microsoft, which are making advances but are not European and therefore may be unfit to address the specific needs of a multilingual Europe.

With notable exceptions, rather than building on the important results and success stories generated by publicly-funded HLT projects, Europe has tended to pursue isolated research activities with a less pervasive impact on the market. However, recent initiatives such as META-NET, Cracking the Language Barrier federation, LT-Innovate and the Connecting Europe Facility programme have attempted to shift this trend by starting to bring the fragmented community together.

Furthermore, the European HLT community acknowledges a lack of coordination between research efforts and the market of HLT applications and services. Recent proposals, originating from this community, have attempted to address these concerns by recommending a concerted European effort on the part of administration, research and industry; thus allowing European HLT SMEs to compete in the global landscape, while at the same time contributing to reduce the technology gap for all European languages.

Language barriers are likely to have significant social and economic consequences

In a multilingual Europe, language barriers are expected to have profound and intertwined social and economic consequences such as: (1) fostering a language divide, (2) hampering workers' mobility, (3) hindering the access to cross-border public services, (4) limiting citizens' engagement and participation in the political process, and (5) creating fragmented markets for cross-border trade and e-commerce, particularly for SMEs.

Regarding the social effects of language, promoting an English monolingual policy in the EU would leave behind 60 % of the European population, with high disparities among countries. In countries such as Hungary, Spain, Portugal and Bulgaria less than 20 % of the population is able to speak English compared to 80 % in the Netherlands. Moving towards a multilingual scenario involving the six most spoken languages (English, German, French, Italian, Spanish, and Polish) would improve the situation, however 15 % of the European population would still be left behind. In 11 European countries (Finland, Greece, Estonia, Lithuania, Slovakia, Czech Republic, Latvia, Romania, Bulgaria, Portugal and Hungary) more than half of the population would be unable to properly communicate using any of those languages.

Furthermore, encouraging policies based on a few languages, besides being unfair to speakers of smaller and minority languages, would create a profound digital barrier, leaving behind less educated and older populations. When analysing the relationship between socio-demographic factors and the ability to speak a foreign language, the results suggest that education is the characteristic most strongly related to speaking a foreign language. On average, leaving formal education after 19 years old increases the chances of speaking English as a foreign language about 19 times more and the chances of speaking one of the six most spoken European languages about nine times more, compared to leaving formal education before 16. Age is also relevant, as younger people (below 30) have, on average, five times more chances of speaking English as a foreign language and four times more chances of speaking one of the six most spoken languages compared to older people. The effect of social status, although important, is less influential than education (high social status increases the chances of speaking a foreign language around three times more compared to low social status). Language policies based on a few selected languages would therefore create a profound language divide.

In the new global environment, language may become a strong barrier against mobility between EU countries. Living in a country without properly speaking the official language imposes a burden over the migrants who are unable to find a job or successfully access basic public services, such as e-

government, health or emergency services. In fact, in the EU, there are about 12 million citizens (2.42 %) who are unable to speak, to a decent level, the official language of the country they live in.

There is a strong influence of language barriers on the mobility of EU citizens within the EU to a different member state. The percentage of EU citizens who have ever moved to a different country within the EU to live and work is only 5.8 %. Language differences are likely to be one of the main barriers hindering workers' mobility within Europe. On average, having low language barriers between two EU countries (i.e., having similar or linguistically related languages) increases more than three times the number of people that decide to move between those countries to live and work. In a scenario where low language barriers exist between all EU countries, the current percentage of 5.8 % would have increased almost three-fold up to 16.1 %. In fact, in 2014 only 1.8 million Europeans (0.26 %) migrated to a different member state, compared to five million people (1.63 %) that moved to a different state within the United States. The internal migration rate in the USA is therefore 6.4 times higher compared to the EU. Lowering language barriers could almost halve the working population mobility rate difference between the EU and the USA.

One of the reasons that could partially explain why language barriers are so important for EU mobility is the lack of public services in the destination country provided in the language of the EU migrant. In a Digital Single Market, with free movement of citizens and goods, public administrations should provide efficient and cross-border citizen services. However, the European Union's internal market is also fragmented regarding e-Government services. Out of the 66 % of public administration portals that offer information in languages different from the country's official language, only 39 % offer information in a language other than English. For example, in the realm of healthcare, a public service where multilingualism plays a relevant role, the Directive 2011/24/EU about patients' rights in cross-border healthcare does not explicitly include the patient's right to be able to communicate in a language they understand when seeking medical care.

Language barriers also affect the construction of a European identity. Citizens and interest groups would only engage in meaningful conversation with decision-makers if it is done in their native language, therefore an important part of the community's social intelligence would be lost. Participation and collaboration in Europe requires that information and data be fully provided in all official languages and that citizens, businesses and civil society are able to address European, national and regional authorities as well as other stakeholders of member states in their native languages.

The Digital Single Market (DSM) is fostering a global European area where firms, workers and citizens can share knowledge, services, products and labour force. In fact, one of the policy goals of the DSM strategy is to improve access for consumers and businesses to online goods and services across Europe. However, language barriers are hindering the achievement of this goal. Only 16 % of European citizens have purchased online from other EU countries in 2015, accounting for 30 % of total e-commerce users in that year, with symptoms of stagnation. About 50 % of European consumers think that they are not prepared to buy in another EU language. It is likely that the language barrier (a hidden barrier for many on-line consumers) is substantially bigger than perceived by consumers because the users never arrive to e-commerce pages in other languages and therefore the consumers are not aware of the problem. People are not aware about what they do not know. Regarding web merchants, the number of companies that are doing electronic sales is increasing although the percentage of companies selling cross-border in the EU, out of the total number of companies selling through electronic channels, is stagnated at around 44 %. Close to 30 % of the companies that sell on-line to individual consumers consider that language is an important barrier.

The average number of cross-border e-shoppers between countries with low language barriers is four times higher compared to countries with high language barriers (i.e., with linguistically distant languages). As a result, the e-commerce in Europe is not working as a truly integrated Digital Single Market. In fact, there is a strong fragmentation into six groups of almost isolated countries. The groups are usually shaped around a big country and most of the countries within the groups share the same or

closely similar languages. Cross-border e-commerce between countries of the different groups is scarce, except when the destination country is one of the biggest economies. Additionally, there is a group that includes most of the countries with smaller languages in Europe which remains isolated and clearly disadvantaged because no European on-line buyers from other countries tend to buy in these countries.

For countries with smaller languages this phenomenon has profound consequences. Take Hungary for example, a country with an isolated language and a low percentage of the population speaking foreign languages. Currently, the people from the rest of the EU shopping on on-line Hungarian websites is negligible. Overcoming language barriers could potentially increase the on-line transactions coming into Hungary from other countries up to 5.9 million, more than twice the current population buying on-line in Hungary (2.2 million).

This effect is particularly negative for SMEs, a crucial pillar of the European economy. Up to 41 % of large companies sell through e-commerce channels domestically and 23 % to other EU countries, compared to 16 % and 7 % for small companies, respectively. One of the reasons for this gap is the language barrier. In fact, the only significant barriers to the number of electronic sales between large and small-and-medium enterprises are the language problems and technical issues.

All in all, the analysis shows that language barriers are likely to have a very profound negative effect on several cross-border social and economic activities in Europe, seriously challenging the **unity** of Europe and the creation of a truly integrated DSM.

HLT are not properly considered in current policies of the EU

When considering multilingualism policies, there is always a trade-off between effectiveness and fairness, between utopia and reality, and between the preservation of cultural heritage and diversity and fostering an effective integrated global market. Multilingual policies then become an uncomfortable topic. On the one hand, preserving multilingualism is strongly rooted in the essence of European values of cultural diversity and is likely to be the only way to create a fair and truly integrated European Union while preserving our culture. On the other hand, it seems that a feasible solution doesn't exist to overcome language barriers. The EC itself has opted for efficiency, moving towards a monolingual regime within their institutions where English has in practice become the only working language.

HLT seem the only feasible way to overcome language barriers while preserving cultural diversity and linguistic rights. Therefore, it should be expected that HLT have a very relevant standing in the policy-agenda of the EU. However, by using text-mining techniques to compare language technology related terms to other trending technology terms, our analysis of over 3.000 technical, political and strategic documents and posts of EU institutions shows disappointing results. Language technology has a low relevance compared to other trending technologies, and the gap is increasing. This is quite surprising if we consider the role that these technologies can have in dealing with the challenge of language barriers preventing a truly integrated EU.

Neither multilingualism nor HLT are properly reflected in current Information and Communication Technologies (ICT) policies in the EU. In fact, the Digital Single Market Strategy of 2015 only makes a brief mention of multilingual services. There are no mentions of either the role of HLT in providing these services or that language is one of the most significant barriers for the EU DSM. The good news is that within the DSM strategy there are two actions that are focused on promoting research on HLT and providing these technologies as a service, although the funding is too scarce to make a difference.

Some EU countries, such as the UK, Ireland and particularly Spain, as well as other multilingual countries outside of Europe, such as India and South Africa, have developed interesting plans regarding multilingualism and technology that could provide useful policy ideas. Eventually, the hope is that industry and research groups will develop more strategic plans that provide valuable insights into multilingual technology as well.

Policy recommendations to effectively draw upon HLT

Multilingualism in Europe is a complex topic involving many stakeholders with intertwined interests in different countries. Therefore, no single policy can tackle the problems described in the previous sections. On the contrary, to truly seize the opportunities of a multilingual Europe we believe that a joint and coordinated action at the European, national and regional levels involving stakeholders from the public sector, civil society, research institutions and industry is required. Therefore, launching a multidisciplinary European Human Language Project including multiple actions is proposed.

We suggest that the EC should be in charge of coordinating the initiative. Not only European institutions, but also national and regional governments should be responsible of creating resources for their languages. Research in Europe should focus on fostering the new deep learning paradigm in HLT, while at the same time providing support for smaller European languages through technology transfer. Talent scarcity and drain brain should be transformed into talent creation and brain gain. Coordination between research and industry should be provided in a seamless, open and effective way through existing European platforms. The public sector should provide their contents and services for all European languages while promoting the growth of the HLT market through public procurement of innovative technology. Mechanisms to facilitate the scaling-up of European innovative HLT companies should be established. Eventually, incentives for firms across Europe to provide their contents, products and services in the different European languages should be provided to create a fully integrated DSM.

To achieve these goals, a set of policy options by using a multi-criteria analysis are proposed and assessed. The policies are structured into five groups: institutional policies, research policies, industry policies, market policies and public services policies.

Institutional policies involve initiatives to adapt current institutional frameworks to draw upon emerging technology trends to better fit the challenges of a multilingual Europe while properly assessing the results. Research policies focus on moving Europe towards the development of the next generation of Language Technologies. Research policies also aim at integrating research and industry, providing Europe with the tools to share resources to effectively compete with other markets. At the same time, these tools will help contribute to the equality of all European citizens in their everyday digital experience regardless of their language. Industry policies foster the creation and growth of competitive European firms while increasing the availability of highly qualified workers. Market policies seek to improve the HLT sector in Europe by raising awareness among European stakeholders of the relevance of these technologies to further increase the demand of their services. There is a specific policy targeting small web merchants so they can benefit from accessing a much bigger market by translating their web shops using HLT. Public service policies intend to create multilingual public services in the European, National, Regional and Local administrations while contributing to increase the innovative HLT sector by using public procurement tools.

The proposed policy options are the following:

- Institutional Policies:
 - Reinforce the role of HLT within the institutional framework of multilingualism-related bodies.
 - Create tools to properly evaluate HLT policies.
- Research Policies:
 - Refocus and strengthen research in LT through the Human Language Project.
 - Promote the European LT Platform for data and services.
 - Bridge the technology gap between European languages.
- Industry Policies:
 - Foster and support the development of investment instruments and accelerator programs targeting HLT start-ups.
 - Increase the availability of qualified personnel in HLT.

- **Market Policies:**
 - Raise awareness of the benefits for companies, public bodies and citizens of the availability of on-line services, contents and products in multiple languages.
 - Promote the automated translation of e-commerce websites of European SMEs.
- **Public Service Policies:**
 - Public procurement of innovative technology and pre-commercial public procurement.
 - Foster the translation of national and regional public websites and documents to other EU languages by using HLT.

1 Introduction

The digital age represents both one of the biggest opportunities for growth and one of the most substantial challenges for the European Union (EU).

Language is at the very basis of this digital world – conversational interfaces, personal assistants and access to information and knowledge. The multilingual nature of Europe and the fact that different languages coexist naturally in European businesses, organisations and schools appear both as a challenge and as an immense opportunity.

For the EU to take advantage of these digital opportunities it must adequately address several challenges, such as configuring a true Digital Single Market (DSM) and providing access to public services and contents regardless of the language spoken by its citizens.

European citizens need to communicate across the language borders of this unavoidable European DSM. Human Language Technologies (HLT) can help overcome this critical barrier while supporting the free and open use of individual languages. In the digital age, communication among people, and between people and machines, as well as unrestricted access to knowledge should be equally accessible for all European citizens regardless of their native language.

However, the HLT European industry is facing serious challenges and non-European top HLT players, such as Google or Microsoft, may be unfit to address the specific needs of a multilingual Europe. In fact, there is a widening technology gap between English and the other official, co-official or non-official EU languages, some of which might already be facing digital extinction.

In this context, our first goal is to describe the current status of HLT in Europe and to quantify the economic, social and linguistic consequences of language barriers in the digital age if no well-designed policy actions are taken.

Our second goal is to assess whether or not multilingualism and HLT are properly reflected in current Information and Communication Technologies (ICT) policies in the EU.

Our final goal is to make a policy analysis to assess the different options available to European policy-makers to ensure language equality in Europe.

The study is structured as follows:

- **Chapter 2, Methodology and resources used**, describes the methodology used in the study.
- **Chapter 3, Synthesis of the research work and findings**, analyses the current status of HLT in Europe, quantifies the socio-economic effect of language barriers in Europe, and describes current multilingualism and HLT policies in Europe.
- **Chapter 4, Policy options**, recommends and assesses different policy options to address the major challenges identified. The assessment is made by using a multi-criteria analysis that takes into consideration different social, economic and political factors.
- Last, **Chapter 5: Conclusions**, summarizes the main findings and makes recommendations for European policy makers based upon the policy assessment.

2 Methodology and resources used

The first methodological resource used was desk research. Given the complexity and extension of the topic and the variety of research questions, diverse sources of information were used such as:

- Reports and brief notes from main public and private stakeholders directly related to the topics. The vision of the industry is mainly provided by LT-Innovate (LT-Innovate, 2012b, 2013, 2016a). The researchers standing is mostly provided by the Cracking the Language Barrier federation (CRACKER & LT_Observatory, 2015). The public vision is reflected in multiple documents from European institutions. We focused on documents about European policies in the digital age and in the DSM, particularly those mentioning multilingualism and language technologies.
- Reports and brief notes from specialized agents such as think-tanks and research groups. The reports related to the META-NET initiative were particularly interesting (Rehm et al., 2014, 2016).
- Academic reports and papers, which tend to take a neutral and objective position while providing quantitative evidence. We have used Google Scholar by searching terms related to the topics covered by the study such as: Digital Single Market, Human Language Technologies, Minority Languages, e-commerce and combinations of those terms.

A systematic research was also performed by using the references included in the aforementioned reports. It was particularly helpful to identify the most relevant public documents.

One of the goals of this study was to support the main conclusions of the qualitative analysis with evidence. To do so we used public datasets, mainly from public European institutions (Eurostat, Eurobarometer) and other international organizations such as the World Bank. We also used private datasets when public information was not available (Civic Consulting, 2011). We used these general datasets to create specific datasets about language barriers between countries and the relationship with other topics such as workers' mobility and cross-border e-commerce. The quantitative analysis was performed by using the open source statistic program R.

One point that was worth analysing was the assessment of the stance of European bodies regarding HLT and multilingualism, particularly relating to the DSM. To do so, big data techniques were used, such as web scrapping and text mining, to collect and analyse about 3.000 documents and posts that were likely to reflect the EU opinion and policies related to these topics. The list of documents is included in Annex 7.6.

After making the qualitative and quantitative analyses, external experts were interviewed to provide additional insights and to assess the key findings of the research. The name and position of the experts are included in Annex 7.1., as well as the names of experts involved in reviewing the first draft of this report.

Eventually, all the information became the basis for the policy analysis to define and assess the different policy options available to European policy-makers by using a multi-criteria assessment matrix. The criteria are detailed in Chapter 4.2.

3 Synthesis of the research work and findings

3.1 Human Language Technologies in multilingual Europe

In this chapter, an overview of the main human language technologies and resources is given and the current situation of LT-related industry and research in Europe is analysed. Also analysed is the existing technology gap between English and other European languages, with an emphasis on smaller and minority languages.

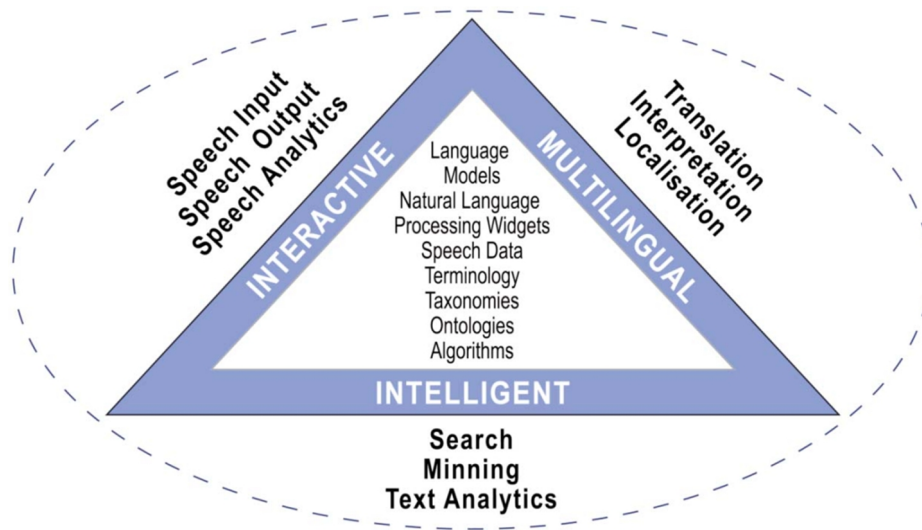
3.1.1 Human Language Technologies: an overview

Human language technologies, simply referred to as language technologies (LTs)¹, are software systems designed to handle human language in all its forms: spoken, written or signed. LTs are found behind many everyday digital products since most of them use language to some extent. Mobile communications, social media, intelligent assistants and speech-based interfaces are transforming the way citizens, companies and public administrations interact in the digital world. LT are the critical enabling technologies for this digital revolution. The impressive progress reached in the last decade allows us to foresee a near future where natural interaction with intelligent applications will be the norm. However, language technologies are, by definition, language-based, and most of the progress so far is related to one language alone: English.

Most HLT-based applications share a large and heterogeneous group of core techniques and tools for language analysis and production, such as tokenisers, part-of-speech taggers, syntactic parsers, information retrieval tools, speech recognition software and speech synthesis engines. Many of these tools depend on specific linguistic datasets or language resources. For example, machine translation systems need large collections of aligned bilingual text. Other tools need specific annotated corpora, treebanks, grammars, lexicons, thesauri, terminologies, dictionaries, ontologies and language models. Below we give an overview of these heterogeneous technologies organized into three main segments as seen in Figure 1:

- **Semantic Technologies**, based on text analytics and information extraction technologies, making it faster and easier to process and search (big) data;
- **Speech Technologies**, allowing a natural interaction based on conversation between people and machines; and
- **Machine Translation**, the ultimate language technology; able to crack the language barrier.

¹ Language technology is a well-established research area with a rich literature. The interested reader can find more information in two good introductory works: (Jurafsky & Martin, 2009) and (Manning & Schütze, 1999)

Figure 1: Language Technologies: tools and resources

Source: Taken from (LT-Innovate, 2013)

3.1.1.1 Semantic Technologies

Semantic technologies are a cover term for all technologies capable of transforming raw text into “intelligent content”. Intelligent content refers to text content that has been analysed with the proper text analysis tools and subsequently enriched with structural and semantic annotations.

For example, the sentence “*Joseph Bech was the Luxembourgish politician that helped set up the European Coal and Steel Community in the early 1950s and a leading architect behind European integration in the later 1950s*”, may be enriched with annotations such as:

Joseph Bech: PERSON (https://en.wikipedia.org/wiki/Joseph_Bech)

Luxembourgish: from Luxemburg PLACE (<https://en.wikipedia.org/wiki/Luxembourg>)

European Coal and Steel Community: ORGANISATION
(https://en.wikipedia.org/wiki/European_Coal_and_Steel_Community)

Early 1950's: TIME

European integration: CONCEPT (https://en.wikipedia.org/wiki/European_integration)

Relationships between the entities may also be defined. For example:

Is_from [Joseph Bech, Luxemburg]

Set_up [Joseph Bech, European Coal and Steel Community]

Leading_architect [Joseph Bech, European integration]

Through using different analysis tools many other annotations could potentially be added: grammatical information, synonyms, ontology relations, sentiment or polarity, etc. Eventually, the original string of characters becomes a structured multidimensional object capable of providing intelligence or semantic interpretation to a suitable application.

Indeed, analysed text is at the core of any intelligent technology or application, such as semantic search, recommendations, reasoning, dialogue management, question answering, opinion mining, text summarization, etc.

Natural Language Processing (NLP) is the discipline traditionally concerned with the process of automatically extracting meaningful information from text. Nearly all current research in NLP is based on statistical machine learning and, more recently, on neural networks or “deep learning.” Most of the earlier systems (roughly until the 1990s) consisted of large sets of rules manually encoded by expert linguists. Some of these rule-based (or symbolic) systems are still used by the industry. As we will see in more detail when discussing machine translation, increase of computational power and access to an ever-growing amount of digital data have boosted the use of statistical approaches that allow extracting, or learning, of linguistic rules automatically through the analysis of large corpora of text.

To give a general idea of the field, we have compiled in Table 1 a non-exhaustive list of the more common NLP tasks along with a short description of each one. Many of them are sub-tasks that can be pipelined to build up larger NLP tasks or applications.

Table 1: Common NLP tasks

NLP Task	Description
<i>Co-reference resolution</i>	Determine which words ("mentions") refer to the same objects ("entities") in a text. E.g. the three underlined mentions in: <u>Angela Dorothea Merkel</u> is a German stateswoman. <u>Merkel</u> has been the Chancellor of Germany since 2005, and <u>she</u> is the leader of the Christian Democratic Union (CDU) since 2000.
<i>Discourse analysis</i>	Identify the discourse structure of connected text.
<i>Morphological analysis</i>	Separate and/or label morphemes with category labels. E.g. <i>bella</i> (IT) <i>singular, feminine</i>
<i>Lemmatisation</i>	Group together the different inflected forms of a word so they can be analysed as a single item, usually the lemma. E.g. <i>bello</i> (<i>bella, belle, belli</i>)
<i>Named entity recognition (NER)</i>	Identify mention of proper names in text and what the type of each such name is (e.g., person, location, organization)
<i>Entity Linking</i>	Determining the identity of entities mentioned in text. For example, the word <i>Washington</i> in “ <i>suburban Washington</i> ” is linked to Wikipedia entry <i>Washington, D.C.</i> , and not to <i>Washington_(state)</i> or <i>George_Washington</i> .
<i>Part-of-speech tagging</i>	Determine the part of speech for each word in the sentence (i.e. noun, verb, adjective, etc.) and possibly the subtypes of those.
<i>Parsing</i>	Analysis of a sentence which identifies constituent parts (e.g., noun phrases, verb groups, etc.).
<i>Semantic role labelling</i>	Detection of the semantic arguments associated with the predicate or verb of a sentence and their classification into their specific roles. For example, in the sentence “Mary sold the book to John”, “to sell” represents the predicate, “Mary” the seller (agent), “the book” represents the goods (theme), and “John” the recipient. Note that the sentence “The book was sold by Mary to John” has a different syntactic form, but the same semantic role.
<i>Relation extraction</i>	Identify the relationships among entities. For example, from the sentence “Barack Obama married Michelle Obama in 1992”, we automatically figure out that “Michelle Obama” “is the wife of” “Barack Obama”.

<i>Sentence segmentation</i>	Find the sentence boundaries in a text.
<i>Topic segmentation and identification</i>	Separate a text into segments, each of which is devoted to a topic, and identify the topic of the segment.
<i>Tokenisation</i>	Break a stream of text up into words, phrases, symbols or other meaningful elements, called tokens, which will then become input for further processing such as parsing or text mining.
<i>Word sense disambiguation</i>	Identify which meaning of a specific word is used in a sentence when the word has multiple meanings.
<i>Truecasing</i>	Determine the proper capitalization of words where such information is unavailable.
<i>Language identification</i>	Determine which natural language a given content is originally provided in.
<i>Terminology extraction</i>	Automatically extract relevant terms from a given corpus of text.
<i>Language checking (spell-checking, grammar correction, authoring support)</i>	Flag misspelled words and constructions containing grammatical errors or not following specific technical guidelines.

Source: Compiled by the authors.

A fair number of open and proprietary online platforms exist that provide text analysis services, among them the Natural Language Toolkit (NLTK) from the University of Pennsylvania, the Cloud NLP API by Google, the LUIS service by Microsoft and the NL Classifier from IBM.

3.1.1.2 Speech technologies

Speech technologies are human-to-computer interfaces providing multimodal, including natural spoken, interaction. They aim at enabling people to communicate with any device using spoken language. Voice is an essential interface in a mobile world and for anyone in need of a hands-free approach to computing or communication tasks. It was not until the late 1990s that companies started to invest in interactive speech systems, which soon became vital for call centres. At the time, speech recognition vendors focused on customer services with basic speech-recognition applications such as voice activated dialling, call routing, etc. Currently, sophisticated applications can accept widely varied and highly complex caller requests, enabling fully automated transactions or customer self-service including, but not limited to, accepting payments and entertainment ticketing, banking transactions or collecting personal information. In fact, nearly every industry segment (communications, financial services, government, healthcare, retail, tourism, etc.) has now implemented automated speech dialogue at some level, from simple call routers to fully automated self-service to even purchase/transaction applications.

Both ways of a dialogue may be covered by machines: speech generation (or Text-to-Speech, TTS) and speech analysis (or Automatic Speech Recognition, ASR). Most current ASR systems are either:

- Speaker independent, i.e., able to recognize the speech patterns of a large group of people and respond to many users.
- Speaker dependent, i.e. able to recognize speech patterns from only one person who has specifically “trained” the system. This technology is seen in commercial dictation packages for medical, legal and business professional transcription.

With respect to Text-to-Speech systems, improvements in this technology, combined with platforms requiring interactivity (such as mobile or gaming), are opening new opportunities for speaking applications. Some notable features are naturalistic voices in many more languages, which are used in education and gaming environments, and interactive access to the web.

Looking into the future, according to eminent voices such as Google CEO Sundar Pichai (Bort, 2016), we are moving from a mobile-first to an Artificial Intelligence (AI)-first world. Spoken LT are part of many AI scenarios that are quickly becoming mainstream, such as the ones listed below that were identified by the Conversational Interaction Technology Innovation Alliance² roadmap. Note that many of these scenarios also incorporate semantic and machine translation technologies, in addition to spoken language technologies:

1. Adaptable conversational interfaces for all:
 - a. Products and services for adaptable interfaces, which will provide augmented reality solutions to local interaction problems on the basis of what they know about you. Your smart personal assistant will transmit data about you to your car, TV set, etc. to ensure that these devices adapt properly.
 - b. Wearable intelligence devices such as smart glasses, smart watches, various types of body monitoring devices, and smart clothing will make some static information media become interactive.
 - c. Automatic interface subtitling will openly help overcome the language barrier by automatically translating spoken and written language for the user.
2. Smart personal assistants:
 - a. Generic personal assistants will act as a smart support service covering all kinds of information, reminders and support services related to life and work, from calendars to available local bargains to TV and film watching management. They will make predictions or recommendations based on personal data and the relation between the user and the real and virtual world. They will filter links to ease access to personal medical agendas, personal travel, banking, and the sensitive management of networks of friends, colleagues and entourage, relevant social media alerts, etc.
 - b. Context-based personal assistants, which will be used for different types of jobs, home management, travelling/tourism, driving vehicles, etc., often closely tied to a particular object or “thing”.
3. Active information access:
 - a. Interactive multimedia search over text, video, image, sound, music, speech, emotion, tonality, value systems, etc. In the long term this can lead to fully conversational search and event alerts including a predictive management of them (disease, accidents, crises, weather, ecology, etc.) using links to knowledge bases, databases, experts, etc.
 - b. Immersive intelligence. Enabling professional communities to create “worlds” in which they can better examine, track and understand a given rich data context. This can lead to e-discovery on an unprecedented scale in areas such as finance, law, pharmaceutical and

² Conversational Interaction Technology Innovation Alliance. (Renals et al., 2015; ‘ROCKIT Project: Roadmap for Conversational Interaction Technologies’, n.d.)

medicine. At the consumer level, it will provide rich multimodal ways to interact with data (3D, avatars, etc.).

4. Communicative robots:

- a. Social and domestic robots will provide interactive and not just physical support. The focus is on “social” where interaction modality is critical. In addition to domestic robots, it could also be argued that developments in cars may be interpreted in terms of development of social robots.
- b. Toy robots. Social robots have enormous potential as intelligent, learning toys. A key trend is the importance of applications aimed at children and young adults, and the importance of “fun” in the use of novel technologies.
- c. Office/workplace support robots.

5. Shared collaboration and creativity

- a. Gaming interfaces. Games have proven to be an important proving ground for many new technologies. We expect this to be true for conversational technologies combined with 3D, augmented reality, haptics and gestures. Furthermore, social interactive games can have added value through enabling players to communicate in different languages and across different cultures.
- b. Support for meetings and collaboration. Online meetings still tend to be rather awkward, but bringing together people with similar interests to share their activities and enable them to easily communicate and work on a common problem, while keeping in mind the existing language barriers among them, could have great potential impact. This might involve videoconferencing with telepresence and integral multilingual support, knowledge monitoring and discreet predictive prompting during the meeting, meeting minute’s management, post-meeting communications or meeting summaries.
- c. Film/video production accelerator deploying video, audio, speech and language processing technologies in the production and postproduction processes.

3.1.1.3 Machine Translation

The idea of using digital computers to translate natural languages can be traced back to the 1950s in the Cold War context. The method at the time was to use bilingual dictionaries in order to do word-by-word translation, but that alone was unable to produce a good translation because of polysemy, wide differences among languages in word order and grammar systems, etc. Ambiguity of human language is a major handicap, both at the lexical level, such as polysemy (e.g., the Spanish word *muñeca* translates into English both as *wrist* and *doll*), and at the syntactic level (e.g., “*The woman saw the car and her husband too*” may have two divergent translations in German: “*Die Frau sah das Auto und ihr Mann auch*” and “*Die Frau sah das Auto und ihren Mann auch*”).

Rule-based Machine Translation (MT) systems were an improvement over word-by-word translation; they were based on grammars written by expert linguists. Rule-based (or knowledge-driven) systems analyse the input text and create an intermediary symbolic representation from which the translation can be generated into the target language. The accuracy of these systems is highly dependent on the availability of extensive monolingual and bilingual lexicons, complete with morphological, syntactic and semantic information, as well as large sets of grammar and translation rules which are carefully designed and tested. It is therefore a long and costly process, potentially endless due to the inherent complexity of human language and the translation process.

In the late 1990s and early 2000s, as computational power and access to digital data increased and became cheaper, interest in statistical approaches to MT started to grow. Statistical, or data-driven MT

systems (SMT), automatically extract translation rules from already translated text, what is called a *parallel corpus*, also referred to as a bilingual corpus, where every sentence, or segment, is aligned to its corresponding translation in the other language. From the aligned segments, SMT systems learn bilingual mappings between words or phrases. Important parallel corpus that have been commonly used to inform SMT systems are the multilingual UN corpus and the Europarl, which contains the proceedings of the European Parliament in 11 European languages. Data-driven MT is advantageous, not only because less human effort is required, but also because it can cover special particularities of the language (e. g., idiomatic expressions) that are overlooked in knowledge-driven systems. On the downside, they have limitations particularly regarding word-order differences and long-distance dependencies. The fact that strengths and weaknesses of knowledge-driven and data-driven systems tend to be complementary is conveniently exploited by hybrid approaches which combine both methodologies.

Still, due in part to the high cost of computing resources, rule-based MT systems dominated the field between 2000 and 2010. There are still companies that offer good rule-based solutions today, but most often they are hybrid solutions. It wasn't until 2010-2012 that Moses (Koehn et al., 2007), a free SMT system funded by the European Commission mostly through projects such as Euromatrix, LetsMT! and others, became the foundation upon which nearly every new commercial SMT system was based. SMT shifted the focus from linguists writing rules, to acquiring aligned corpora which are necessary to train SMT systems.

Translation is by far the most service-intensive segment in LT. In the professional market, a quiet but pronounced shift from "human translation only" to "machine translation with post-editing" is taking place in the translation agencies, some of whom are also actively developing their own engines using open-source software like *Moses*. The cost of human translation without automation is simply unsustainable for most applications. Current machine translation systems generally translate with enough quality for comprehension, but not for publication. For high quality translation, human revision and post-edition is still a must.

Over the past few years, it seemed as if SMT had reached a kind of quality plateau, until a "new old" paradigm, coming from the 1980s Artificial Intelligence field, re-emerged in 2014, initially for image processing: Neural Networks. An artificial neural network is a model inspired by the functional aspects and structure of the human brain, which is used to estimate or approximate functions that can depend on a large number of inputs, generally unknown. The big breakthrough brought by Neural Machine Translation (NMT) over the last two years is that it learns to map a whole sentence, from source to target, all at once, instead of word-by-word, phrase-by-phrase, or piece-by-piece, as with earlier approaches³. This eliminates the problems of long-distance dependencies and word-order variations because the system learns whole sentences at once. NMT has shortcomings as well. Neural networks require a lot of training data, in the order of one million sentence pairs, and there's still no good solution to translating rare or unseen words, although there have been a few proposals on how to address this problem, such as character-based modelling. The other major shortcoming is that very powerful hardware is required in order to train large neural networks efficiently. If SMT already requires a lot of memory to store translation models and CPUs with multiple cores in order to parallelize training, NMT requires higher-end GPUs (i.e. graphics processing units) for training.

Many talks and posters at MT conferences are currently dedicated to advancement and progress in NMT, and big players, such as Google and Microsoft, are both working on ways to use NMT in their translation products, with a special interest in how NMT can significantly improve fluency in

³ Neural Machine Translation is the Next Big Thing. Blogpost by Dave Landan. May 2016

translation between Asian and European languages. All the signs show that we can expect major advances in MT in the coming years. As a recent proof of this, the following piece made news on 27 September 2016: “Google says its new AI-powered translation tool scores nearly identical to human translators” (QUARTZ, 2016), explaining that Google’s research Neural Network-based MT system was already at production scale, at least for an initial set of language pairs. European companies are also moving toward production their NMT systems that show significant improvement in translation quality for both major languages (Systran) and smaller EU languages (Tilde).

3.1.1.4 Sign Language Technologies

Sign languages (SLs) are linguistic systems that instead of using oral utterances to convey meaning, use iconic expressions by means of hands, arms, body and facial gestures. A significant difference between sign and oral languages is the way the former use three-dimensional space with a grammatical function, compared to the linear nature of oral utterances.

There can be more than one signed language in a country, just as for oral languages. For example, there exist two sign languages in Belgium (French Belgian Sign Language and Flemish Sign Language) and in Spain, too (Spanish Sign Language and Catalan Sign Language). Also, there are different sign languages in countries that have the same spoken language, such as in the UK and Ireland. This is due to historical developments that are different to the ones experienced for spoken languages.

Similarly to spoken language technologies, SL technologies try to cover both ways of a dialogue: SL generation and SL recognition. A specific field involving both is called SL automatic translation.

- SL generation is a special form of multimodal natural language generation that uses multiple linguistic output channels. It has applications for the generation of gesture animation and other communication signals that are not easily encoded as text strings. Signing avatars can be used to convey SL generation output.
- SL recognition remains a very challenging task in the field of computer vision and human computer interaction.
- Automated SL translation aims at translating an oral (or written) language into a SL, or vice-versa, e.g. English into American Sign Language

Specific sign language technologies involve even bigger challenges than standard LTs, however they can have a dramatic impact in how a deaf person has access to knowledge, information and services.

3.1.2 Linguistic resources and data

3.1.2.1 Corpora

Obviously, text (in digital format) is essential to language technologies. A large collection of digital texts assembled together for a purpose, or simply collected opportunistically, is conventionally called a corpus (plural: corpora). Statistical (also known as machine learning) systems are trained on corpora. Corpora may consist of raw text but they may also contain linguistic annotations associated with that text. Typical examples of linguistic annotations are part-of-speech tags and lemma (or the base form of a word). Figure 2 compares an untagged version of a fragment of the Brown corpus to a version annotated with part-of-speech tags (involving morphosyntactic information).

Figure 2: Untagged version of a fragment of the Brown corpus compared to a version annotated with part-of-speech tags

Brown Corpus Sample (untagged)	Brown Corpus Sample (tagged)
A01 0010 The Fulton County Grand Jury said Friday an investigation	A01_FO 0010_MC The_AT Fulton_NP1 County_NN1 Grand_JJ Jury_NN1 said_VVD Friday_NPD1 an_AT1 investigation_NN1
A01 0020 of Atlanta's recent primary election produced "no evidence" that	A01_FO 0020_MC of_IO Atlanta_93 's_03 recent_JJ primary_JJ election_NN1 produced_VVD " " no_AT evidence_NN1 " " that_CST
A01 0030 any irregularities took place. The jury further said in term-end	A01_FO 0030_MC any_DD irregularities_NN2 took_VVD place_NN1 ._, The_AT jury_NN1 further_RRR said_VVD in_IL term-end_NN1
A01 0040 presentments that the City Executive Committee, which had over-all	A01_FO 0040_MC presentments_NN2 that_CST the_AT City_NN1 Executive_NN1 Committee_NN1 ,_, which_DDQ had_VHD over-all_RR
A01 0050 charge of the election, "deserves the praise and thanks of the	A01_FO 0050_MC charge_NN1 of_IO the_AT election_NN1 ,_, " " deserves_VVZ the_AT praise_NN1 and_CC thanks_NN2 of_IO the_AT
A01 0060 City of Atlanta" for the manner in which the election was conducted.	A01_FO 0060_MC City_NN1 of_IO Atlanta_NP1 " " for_IF the_AT manner_NN1 in_IL which_DDQ the_AT election_NN1 was_VBDZ conducted_VVN ._,

Source: Taken from the Brown Corpus

Some corpora have annotations of further structured levels of linguistic analysis. In particular, a number of smaller corpora may be syntactically parsed. Such corpora are usually called syntactic treebanks. The difficulty of ensuring that the entire corpus is completely and consistently annotated, i.e., manually reviewed, means that these corpora are usually smaller, containing around one to three million words. Other levels of linguistic structured analysis are possible, including annotations for morphology, semantics and pragmatics. Figure 3 shows an annotated sentence of the Penn TreeBank, the most popular syntactic annotated corpus for English, which is regularly used to train statistical parsing tools.

Figure 3: Annotated sentence of the Penn TreeBank

```
((S (NP-SBJ-I Jones)
  (VP followed)
  (NP him)
  (PP-DIR into
    (NP the front room) )
  (S-ADV (NP-SBJ *-1)
    (VP closing
      (NP the door)
      (PP behind
        (NP him) ) ) ) )
.))
```

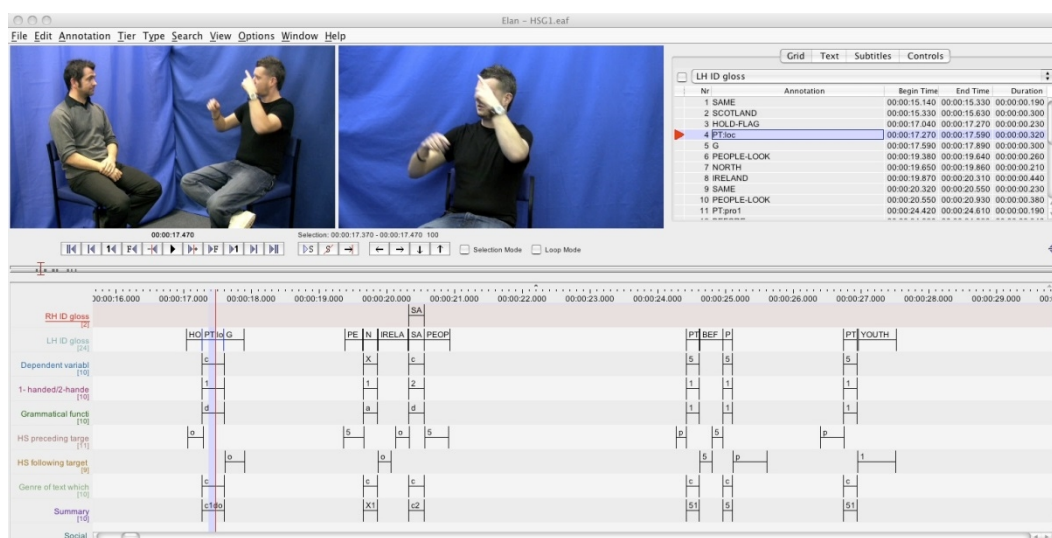
Source: Taken from Penn TreeBank

Corpora can also take the form of audio recordings and audio-visual data. Speech technologies, for example, make use of these kinds of resources. Annotation of multimedia corpora include time-coded transcriptions and linguistic annotations on the transcriptions.

Sign language corpora, on their part, need to record the three-dimensional nature of sign languages, which makes annotation of these languages difficult and costly. Consequently, sign languages tend to be very resource-poor. For most of European SLs there are no corpora at all. The majority of existing sign language corpora are focused on video annotation such as the NCSLGR Corpus (National Center

for Sign Language and Gesture Resources) and the BSL (British Sign Language Project), which uses the ELAN tool for the creation of complex annotations on video and audio resources. In the literature analysed, there were some attempts found to create parallel textual SL corpora such as the German RWTH-PHOENIX-Weather corpus. But, in general, there is a lack of multilingual large parallel corpora for sign languages. This represents a significant obstacle for sign language research and, particularly, for Sign Language MT (Efthimiou et al., 2009).

Figure 4: Screen shot of an ELAN file



Source: Taken from British Sign Language Corpus Project

3.1.2.2 Lexicons and ontologies

Together with annotated corpora, dictionaries and ontologies are other valuable linguistic resources. Broadly speaking, a digital dictionary (or lexicon) is a list of entries (usually single words or multiword expressions) optionally enriched with further information. A full form lexicon is a comprehensive lexical database that contains all inflected, declined and conjugated forms of a language. Unlike an ordinary dictionary that lists only the canonical forms (base lexemes), such as *eat*, a full-form lexicon includes all inflected forms such as *eating*, *eaten* and *ate*. Morphologically rich languages like Spanish can have hundreds of inflected forms for each verb, possibly including clitics. As with corpora, dictionaries can be manually annotated. For example, a lexicon with *polar* annotations (e.g., the verb *improve* is marked positive, while *decay* is negative) can be used for sentiment or opinion analysis.

Ontologies are much more structured than dictionaries. Entries in an ontology (usually called entities) are hierarchically organized and form classes and subclasses. They also have properties and connections among them. A domain ontology represents concepts which belong to a specific domain of knowledge, while an upper ontology (also known as root or foundation ontology) is a model of the common objects that are generally applicable across a wide range of domain ontologies. It usually employs a core glossary that contains the terms and associated object descriptions as they are used in various relevant domain sets. WordNet is a special case of linguistic ontology, where words are grouped into sets of synonyms (synsets), providing a number of relations among these synonym sets or their members themselves. It is a very useful resource for language technology applications. Although the English WordNet is by far the most complete, there exist wordnets for more than 200 languages.

3.1.2.3 Linguistic resources and data in machine learning

Language resources are needed to build HLT tools, both rule-based and machine-learning systems. Machine learning on manually annotated data is called “supervised” training, while learning on raw data is called “unsupervised”. Manual annotations are costly and there is a growing trend towards unsupervised training using ever larger collections of unannotated data. However, to ensure a minimum quality standard, most HLT tasks require a certain degree of supervised training, complemented by unsupervised training on very large corpora. This means that two things are essential in order to build HLT tools for a given language: annotated resources for that language (corpora, lexicons, ontologies) and access to large amounts of data in that language.

Presently, we are witnessing a notable resurgence of Artificial Intelligence (AI) and related fields, such as HLT. But ‘AI is only as good as the data it crunches’ (Vanian, 2016). It is not by chance that this trend is led by big companies like Google, Facebook and Microsoft, who use AI-related techniques to train computers to recognize objects in photos and understand human language. These companies are able to train their computers to perform these difficult feats mostly because they have the enormous quantities of data that are required. When it comes to competition, it is the data these companies possess that is more important than the actual AI software tools they use and release to the public. In fact, according to David Kenny, general manager of IBM’s Watson AI-service, only 20 % of the world’s information is stored on the internet, with the other 80 % being privately held within companies and organizations⁴.

As Clive Humby, UK mathematician and architect of Tesco’s Clubcard, stated back in 2006: “Data is the new oil. It is valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc. to create a valuable entity that drives profitable activity; so must data be broken down, analysed for it to have value.”⁵ And HLT has the tools necessary for the analysis process.

Most improvements in HLT rely particularly on the ability to access and maintain ever larger and more finely tuned linguistic data. Lack of access to that data will constrain the technological development of LT. Acquiring and using it may rely on cooperation between the LT industry and the different constituencies that own, need and use it. Collaboration between the industry and data owners will be important. On top of that, regulation of the use of such data should be made much more open and core language resources (annotated corpora, lexicons and ontologies) should be standardised and shared in an open environment.

3.1.3 Technology gap between English and the other languages

Between 2010 and 2012, a systematic survey of the linguistic particularities of all European languages and an up-to-date status of LT support for them was addressed by the META-NET White Paper Series “Europe’s Languages in the Digital Age”⁶. The survey, prepared by more than 200 experts and documented in 31 volumes, assessed language technology support for each language in four different areas: automatic translation, speech interaction, text analysis and the availability of language resources.

The 30 European languages (including all 24 official EU languages) addressed by this survey were: Basque, Bulgarian, Catalan, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French,

⁴ Oral communication in panel on Artificial Intelligence held in Aspen, Colo. during Fortune’s annual Brainstorm Tech conference (11-13 July 2016).

⁵ First expressed at the ANA Senior Marketer’s Summit 2006 at Kellogg School of Management.

⁶ <http://www.meta-net.eu/whitepapers>

Galician, German, Greek, Hungarian, Icelandic, Irish, Italian, Latvian, Lithuanian, Maltese, Norwegian, Polish, Portuguese, Romanian, Serbian, Slovak, Slovene, Spanish and Swedish.

For each language, support to language technology was categorised using a five-point scale: 1 excellent support; 2 good support; 3 moderate support; 4 fragmentary support; 5 weak or no support, according to the following key criteria:

- **Machine Translation:** quality of existing technologies, number of language pairs covered, coverage of linguistic phenomena and domains, quality and size of parallel corpora, amount and variety of applications.
- **Speech Processing:** quality of existing speech recognition and synthesis technologies, coverage of domains, number and size of existing corpora, amount and variety of available applications.
- **Text Analytics:** quality and coverage of existing technologies (morphology, syntax, semantics), coverage of linguistic phenomena and domains, amount and variety of available applications, quality and size of (annotated) corpora, quality and coverage of lexical resources and grammars.
- **Resources:** quality/size of text, speech and parallel corpora, quality/coverage of lexical resources and grammars.

In 2014, an extension of the survey (Rehm et al., 2014) was carried on in order to address regional and smaller languages that were left out in the 2012 survey. In that year, the White Paper for Welsh was also published. In the extended set, those languages not listed in (Ethnologue, 2013), which had less than 100 000 speakers, and languages not originated in Europe were excluded. The total list of languages included in the cross-comparison appears in the Table 2. Languages which were not addressed in the 2012 survey (shown in italics in the table below), do not have a specific White paper study.

Table 2: Languages included in the updated cross-language comparison

Language	Speakers	White Paper (reference)
<i>Albanian</i>	7 436 990	
<i>Asturian</i>	110 000	
Basque	657 872	Hernández et al, 2012
<i>Bosnian</i>	2 216 000	
<i>Breton</i>	225 000	
Bulgarian	6 795 150	Blagoeva et al, 2012
Catalan	7 220 420	Moreno et al, 2012
Croatian	5 533 890	Tadić et al, 2012
Czech	9 469 340	Bojar et al, 2012
Danish	5 592 490	Pedersen et al, 2012
Dutch	22 984 690	Odijk, 2012
English	334 800 758	Ananiadou et al, 2012
Estonian	1 078 400	Liin et al, 2012
Finnish	4 994 490	Koskeniemi et al, 2012
French	68 458 600	Mariani et al, 2012
<i>Frisian</i>	467 000	
<i>Friulian</i>	300 000	
Galician	3 185 000	García-Mateo and Arza, 2012
German	83 812 810	Burchardt et al, 2012
Greek	13 068 650	Gavrilidou et al, 2012
<i>Hebrew</i>	5 302 770	
Hungarian	12 319 330	Simon et al, 2012
Icelandic	243 840	Rögnvaldsson et al, 2012
Irish	106 210	Judge et al, 2012
Italian	61 068 677	Calzolari et al, 2012
Latvian	1 472 650	Skadiņa et al, 2012

Language	Speakers	White Paper (reference)
<i>Limburgish</i>	1 300 000	
Lithuanian	3 130 970	Vaišnien and Zabarskaitė, 2012
<i>Luxembourgish</i>	320 710	
<i>Macedonian</i>	1 710 670	
Maltese	429 000	Rosner and Joachimsen, 2012
Norwegian	4 741 780	Smedt et al, 2012a; Smedt et al, 2012b
<i>Occitan</i>	2 048 310	
Polish	39 042 570	Miłkowski, 2012
Portuguese	202 468 100	Branco et al, 2012
Romanian	23 623 890	Trandabăţ et al, 2012
<i>Romany</i>	3 017 920	
<i>Scots</i>	100 000	
Serbian	9 262 890	Vitas et al, 2012
Slovak	5 007 650	Šimková et al, 2012
Slovene	1 906 630	Krek, 2012
Spanish	405 638 110	Melero et al, 2012
Swedish	8 381 829	Borin et al, 2012
<i>Turkish</i>	50 733 420	
<i>Vlax Romani</i>	540 780	
Welsh	536 890	Evas, 2014
<i>Yiddish</i>	1 510 430	

Source: Rehm et al., 2014

Table 3 shows the level of support for each of the 30 languages included in the 2012 comprehensive study on the four LT areas: machine translation; speech processing; text analytics and language resources⁷.

Table 3: State of LT support for 30 European languages in four different areas

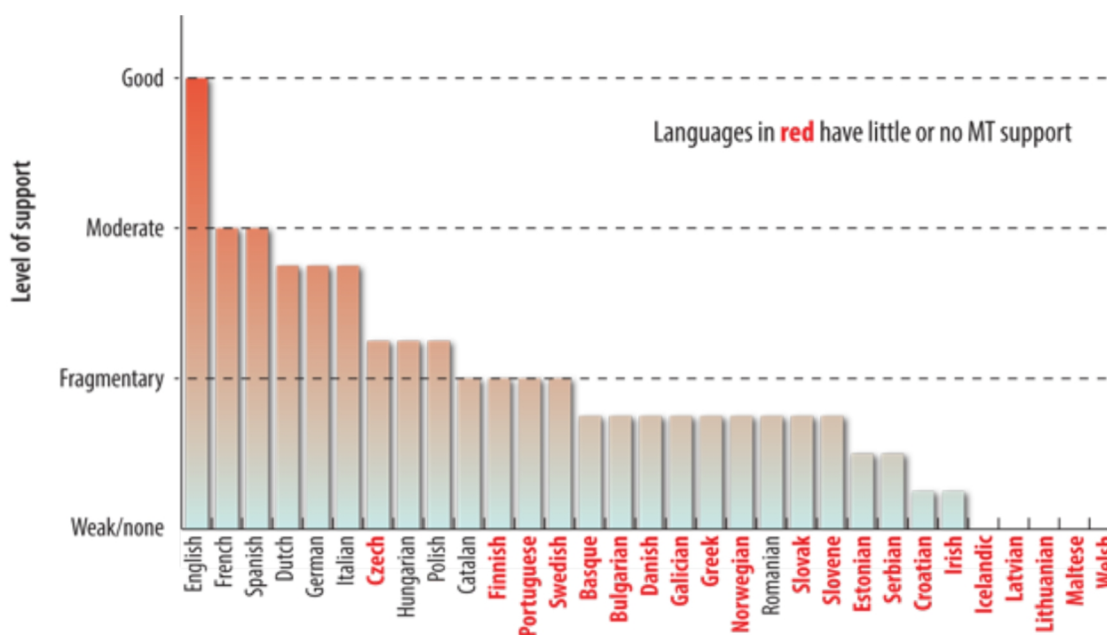
Technology	Good Support	Moderate Support	Fragmentary Support	Weak/no Support
Machine Translation	English	French, Spanish	Catalan, Dutch, German, Hungarian, Italian, Polish, Romanian	Basque, Bulgarian, Croatian, Czech, Danish, Estonian, Finnish, Galician, Greek, Icelandic, Irish, Latvian, Lithuanian, Maltese, Norwegian, Portuguese, Serbian, Slovak, Slovene, Swedish, Welsh
Speech	English	Czech, Dutch, Finnish, French, German, Italian, Portuguese, Spanish	Basque, Bulgarian, Catalan, Danish, Estonian, Galician, Greek, Hungarian, Irish, Norwegian, Polish, Serbian, Slovak, Slovene, Swedish	Croatian, Icelandic, Latvian, Lithuanian, Maltese, Romanian, Welsh
Text Analysis	English	Dutch, French, German, Italian, Spanish	Basque, Bulgarian, Catalan, Czech, Danish, Finnish, Galician, Greek, Hungarian, Norwegian, Polish, Portuguese, Romanian, Slovak, Slovene, Swedish	Croatian, Estonian, Icelandic, Irish, Latvian, Lithuanian, Maltese, Serbian, Welsh
Language Resources	English	Czech, Dutch, French, German, Hungarian, Italian, Polish, Spanish, Swedish	Basque, Bulgarian, Catalan, Croatian, Danish, Estonian, Finnish, Galician, Greek, Norwegian, Portuguese, Romanian, Serbian, Slovak, Slovene	Icelandic, Irish, Latvian, Lithuanian, Maltese, Welsh

Source: Table extracted from the results of the META-NET White Paper series

⁷ Note that all 17 languages included in the extended 2014 study would be placed in the last column (Weak or no support).

Even though more fine-grained analyses are needed, the cross-comparison demonstrates that there are dramatic differences in LT support among the European languages. While there are good-quality software and resources available for a few languages and certain application areas, other (usually smaller) languages have substantial gaps. Digital support for 21 of the 30 languages investigated is “non-existent” or “weak” at best. Figure 5 shows in a more graphical way the level of support of machine translation for each language.

Figure 5: Level of support of MT by language

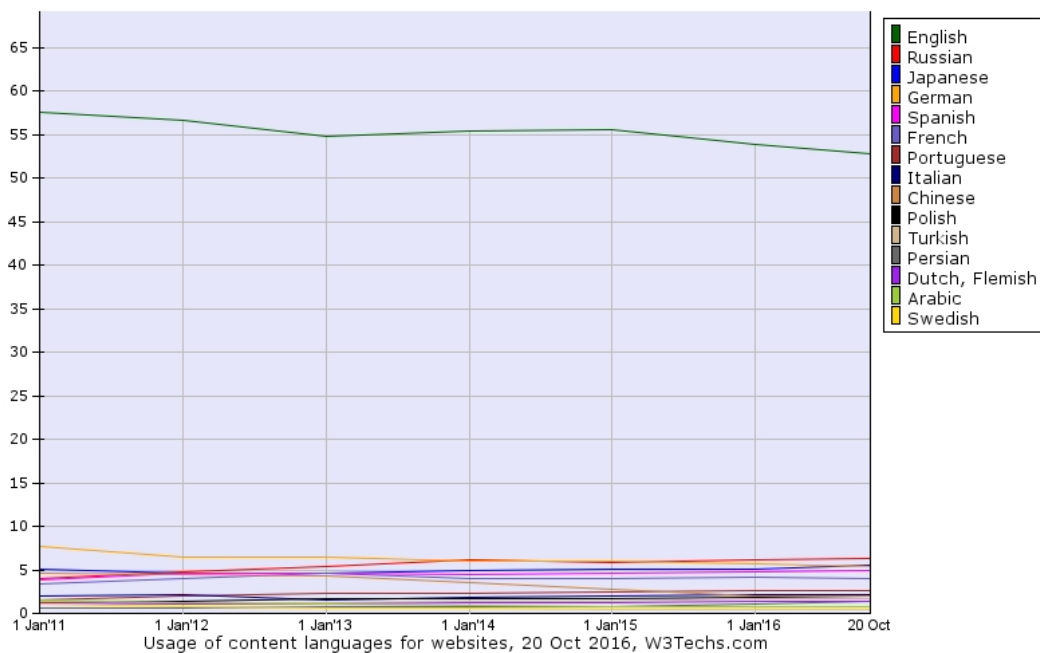


Source: Taken from Language as a Data Type – Strategic Research and Innovation Agenda for the Multilingual Digital Single Market (Version 0.9) – July 2016

The press release that accompanied the initial publication of the Series, with the title *At least 21 European Languages in Danger of Digital Extinction*, was circulated on the European Day of Languages 2012 (26 September 2012) and generated more than 600 mentions internationally (newspapers, blogs, radio and television interviews etc.) (Rehm et al., 2014). This shows that Europe is very passionate and concerned about its languages and that it may be also very interested in the idea of establishing a solid LT base for overcoming language barriers. In several member states like Latvia and Lithuania the study helped to start targeted projects and national programs to improve technology support of the national language.

As shown by Table 3, when compared to the technological development of English, all languages without exception fall behind. It comes as no surprise, considering the universal predominance of English at all levels. To illustrate this fact, see Figure 6, showing language usage in websites, provided by the W3Techs organisation⁸. More specifically, it shows the historical trends in the usage of content languages between January 2011 and October 2016.

⁸ https://w3techs.com/technologies/overview/content_language/all

Figure 6: Language usage in websites (percentage of websites)

Source: Taken from W3Techs

Two observations emerge: the huge gap between the use of English and the rest of big languages (only languages with more than 1 % usage are shown); and the slow reduction of this gap through time, confirming that, hopefully, the future is likely to be multilingual. According to (LT-Innovate, 2013, page 12), ‘the era of the Lingua Franca is over. Interacting across the many languages of the digital world is no longer optional’.

As it turns out, the internet ecosystem is dominated by non-European companies, mainly from the USA. According to the ranking elaborated by the company Alexa (2016), there is not any website from an EU country among the top 20 most visited websites worldwide. The original language of half of them is English, as they are developed by American companies, while the rest have been originally created in Chinese, Russian or Japanese. These websites are the main gate to many digital resources for citizens and companies (communication tools, search engines, e-commerce platforms, social networks, digital contents, etc.) and can be considered a good indicator of how foreign companies deal with European multilingualism in the internet ecosystem. By localising their products and services to the languages used in the markets where they are present, those companies become intensive producers and consumers of HLT. The HLT used by internet giant companies can be developed in-house or, most frequently, obtained through the acquisition of specialised firms. The analysis of the acquisitions made by four of the main internet-related companies (Google, Facebook, Apple and Microsoft) shows an intense buying activity. These companies have acquired more than 20 HLT firms, most of them American, in the last ten years⁹ to support multilingualism in their products and services. This level of acquisitions, which is analysed in detail in chapter 3.1.4, made by only a few number of Internet agents, reflects the strategic role that HLT firms are playing in the internet and the predominance of American providers. These providers benefit from the synergy created among the Internet giants since they belong



⁹ Estimation based on information retrieved from companies’ websites and Wikipedia.

to the same start-up ecosystem, sharing physical locations (for instance Silicon Valley) and, usually, investors.

The dominance of the internet ecosystem by a limited number of players has negative consequences regarding smaller languages, as the localisation of Internet services and contents to smaller languages can be economically inefficient, the companies must focus on maximising their benefits in an increasingly competitive environment. The reduced market size for smaller languages, the cost of localising services and contents and the limited existence of parallel corpora hinder the automation of the translation process, thus minimising the interest of internet companies to offer their services in those languages. To show this phenomenon we have analysed the languages supported by 12 internet services provided by some of the most relevant internet companies, which are in relation to the top 20 most visited websites accessed from the EU. We have selected the languages included in the Eurobarometer survey about Europeans and their languages (European Commission, 2014a), which covers the most spoken languages by European citizens and migrants. The result of the analysis is shown in Figure 7.

Figure 7: Languages supported in selected Internet services

Languages	Google Translate	Google Developer Console	Google	Youtube (interface)	Facebook	Yahoo (web)	Amazon (Kindle Direct Publishing)	Wikipedia	Twitter	Live (outlook)	Bing	Instagram
Arabic												
Basque												
Bulgarian												
Catalan												
Croatian												
Czech												
Danish												
Dutch												
English												
Estonian												
Finnish												
French												
Galician												
German												
Greek												
Hindi												
Hungarian												
Irish												
Italian												
Japanese												
Korean												
Latvian												
Lithuanian												
Luxembourgish												
Maltese												
Polish												
Portuguese												
Romanian												
Russian												
Scottish Gaelic												
Slovak												
Slovenian												
Spanish												
Swedish												
Turkish												
Urdu												
Welsh												

 Not supported
  Supported

Source: Compiled by the authors based in information obtained from websites

The smaller languages (e.g. Irish, Luxembourgish, Maltese, Scottish Gaelic and Welsh) are the least supported in the internet services analysed. In fact, according to the META-NET analysis of the survey results presented above, one key feature in regards to the technology gap is that the number of speakers of a certain language seems to correlate with the amount and quality of technologies available for that language. For companies there is simply no sustainable business case, which is why they refrain from

investing in the development of sophisticated language technologies for a language that is only spoken by a small or very small number of speakers.

While accessibility to internet services, in terms of language availability, seems to be correlated with the number of speakers of a certain language, the affordability of the language technologies that allow this access does not directly affect European citizens. The costs of the language technologies are assumed by the internet services providers, which offer these services either for free or by subscription. Thus, the affordability of language technologies is not a problem for citizens and big companies that can finance their acquisition, but it is a problem for SMEs that want to translate their online services and cannot assume the cost of doing so.

Despite the lack of technological support and available services in languages other than the major ones, regional and smaller languages are increasingly being used in the internet, especially in social media channels. This emerging use of smaller languages is likely to become a major factor pushing forward the demand for language-based technologies in the near future. The Digital Language Diversity Project¹⁰ is currently reviewing the community-driven initiatives that try to make the internet more linguistically diverse.

In order to bridge the technology gap, policies should focus on fostering technology development for European languages other than English, particularly the smaller ones or less-resourced ones, and also on language preservation through digital means. Not all countries have the required expertise or human resources necessary to provide the technology support for their languages. For example, in Iceland there is not a single position in LT at any Icelandic university or college and there is only one company that works in this area (Helgadóttir & Rögnvaldsson, 2013). This is why it is important to establish techniques, methods and instruments for research and knowledge transfer so that countries such as Iceland can benefit as much as possible, for their own language, from the research carried out in other countries for smaller languages. Bootstrapping the set of core language technologies and resources for all languages spoken in Europe is not a matter of a few countries joining forces but a challenge on Europe as a whole that must be addressed accordingly to avoid digital exclusion and secure future business development.

3.1.4 LT-related research and industry in Europe

Europe has a strong scientific base in language engineering and technology, and there is no shortage of innovative new entrants. As exposed by the LT-Innovate (2012) report, many market leaders in the LT industry have deep European roots: SAP is a market leader in analytics applications; Autonomy (a UK firm acquired by HP) is a pioneer in the enterprise search segment using LT; UK-based SDL is a market leader in translation software and Systran (FR) is a leading pioneer in machine translation. A significant proportion of current speech solutions are also based on technology developed first in Europe, although now owned by Nuance in the USA.

Europe has a long-standing R&D tradition with over 800 centres performing scientific and technological research on European languages (META technology council, 2012). The European Union has funded big projects such as EuroMatrix that have produced basic research and resources for establishing high quality language technology solutions for several European languages. European research in the area of language technology has already achieved a number of outstanding successes. For example, since 2013, the translation services of the EU have used the Moses open source machine translation software, which has been mainly developed in European research projects. In addition, national funding used to

¹⁰ <http://www.dldp.eu/>

have a huge impact. For example, the Verbmobil project, funded by the German Ministry of Education and Research between 1993 and 2000, pushed Germany to the top position in the world in terms of speech translation research for a time.

The result of this financial effort is uneven. On the positive side, it has sustained research in universities and research centres, has forged alliances and cross-border scientific communities and has produced innovations and widely-used platforms, such as MT engine Moses. However, on the negative side it has failed, for the most part, to stimulate the European industry to invest in HLT. Thus, rather than building on the important results and success stories generated by these projects, Europe has tended to pursue isolated research activities with a less pervasive impact on the market.

A notable exception to this situation is the case of the already mentioned Moses engine. Research by TAUS (Achim Ruopp & Jaap van der Meer, 2015) 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 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 dollars or about 20 % of the entire MT solutions market.

In 2010, the large-scale META-NET¹¹ initiative (a European network of excellence supported in its first phase by four EU projects), started to bring the fragmented community together and to assemble researchers from different subfields and other related scientific fields (humanities, psychology, social sciences etc.), universities, research centres, language communities, national language institutions, smaller and medium companies as well as large enterprises, officials, administrators and politicians. By now it has more than 650 members in more than 50 countries. META-NET's vision and planning process has involved more than 300 companies, of which more than 200 have already joined the network. With the aim of guiding the European LT community in its goal of fulfilling the technology demands of a multilingual European society and to turn these needs and emerging business opportunities into competitive advantages, the META-NET technology council produced in 2012 the Strategic Research Agenda for Multilingual Europe 2020 (Georg Rehm & Hans Uszkoreit, 2012).

An important contribution of META-NET is META-SHARE, the open resource exchange infrastructure that provides access to thousands of language resources and technologies, and, together with other European initiatives such as FLReNet¹² and CLARIN¹³, promotes a culture of sharing resources. The goal of these initiatives is to have a business-friendly framework to stimulate commercial use of resources, based on a sound licensing facility. META-SHARE is not limited to data. Instead, it may be considered an international hub of resources and technologies for speech and language services from industries and communities, including evaluation protocols and collaborative workbenches. The accumulation and sharing of resources and tools in a single place is intended to lower the R&D costs for new applications in new language resource domains. Innovation in LT crucially depends on language resources but currently there are not enough available resources to satisfy the needs of all languages,

¹¹ www.meta-net.eu

¹² www.flarenet.eu

¹³ www.clarin.eu

quantitatively and qualitatively. Language resources should be produced and made available for every language, every register, and every domain to guarantee full coverage and high quality.

Since the beginning of 2015, new LT projects have been launched by the European Commission and 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 consists of 11 organisations and more than 20 projects. Together with the project LT_Observatory, CRACKER produced in 2015 the “Strategic Research and Innovation Agenda for the Digital Single Market”, and more recently, a follow-up version of the Strategic Agenda, with the name “Language as Data Type and Key Challenge for Big Data”, published in July 2016.

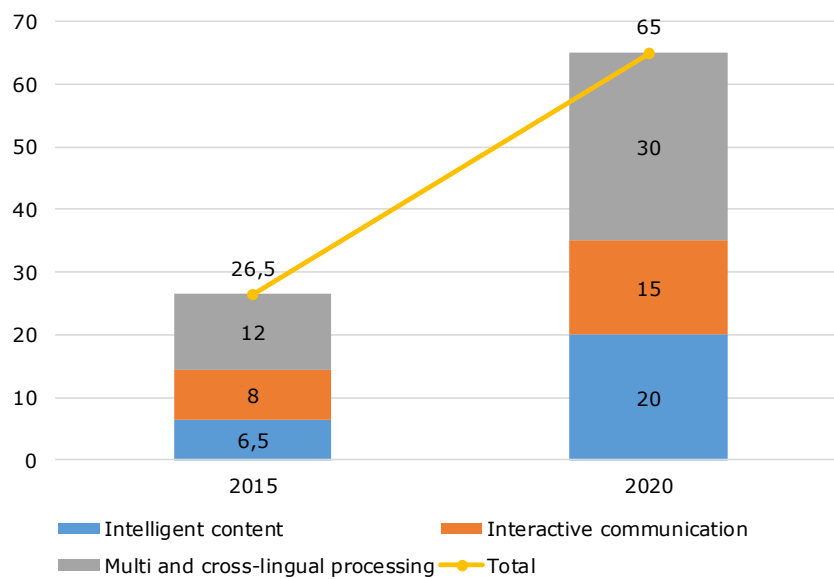
In parallel to the research and innovation-oriented activities funded through FP7 and Horizon 2020, the EC is supporting the deployment of LT within the Connecting Europe Facility programme (CEF). The Machine Translation service, CEF Automated Translation (eTranslation)¹⁴, is built on an existing machine translation system, MT@EC, developed by the EC (DG Translation) and based 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 European institutions and bodies, online services funded or supported by the EU and public administrations in the EU countries, Norway and Iceland. A closer collaboration between CEF eTranslation and the European language technology community has been established through the service contract ELRC (European Language Resource Coordination)¹⁵, which was awarded in April 2015. This collaboration especially takes into regard collaboration via the systematic and coordinated collection and exploitation of language resources in all CEF participating countries, as well as service contracts launched in December 2016.

European R&D has produced a steady stream of small LT-based companies. 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 “Status and Potential of the European Language Technology Markets” in 2013 (LT-Innovate, 2013) and an “Innovation Agenda & Manifesto” in 2014 (LT-Innovate, 2016a, 2016b).

According to these reports, the worldwide LT market - which includes software and services related to intelligent content, interactive communication and multi and cross-lingual processing - reached 19.2 billion euro in 2011. A recent update of the worldwide LT market forecast (LT-Innovate, 2016b) estimates that the revenues will grow from 26.5 billion euro in 2015 to 65 billion euro in 2020, which represents an annual average growth rate of 19.7 % in such a period.

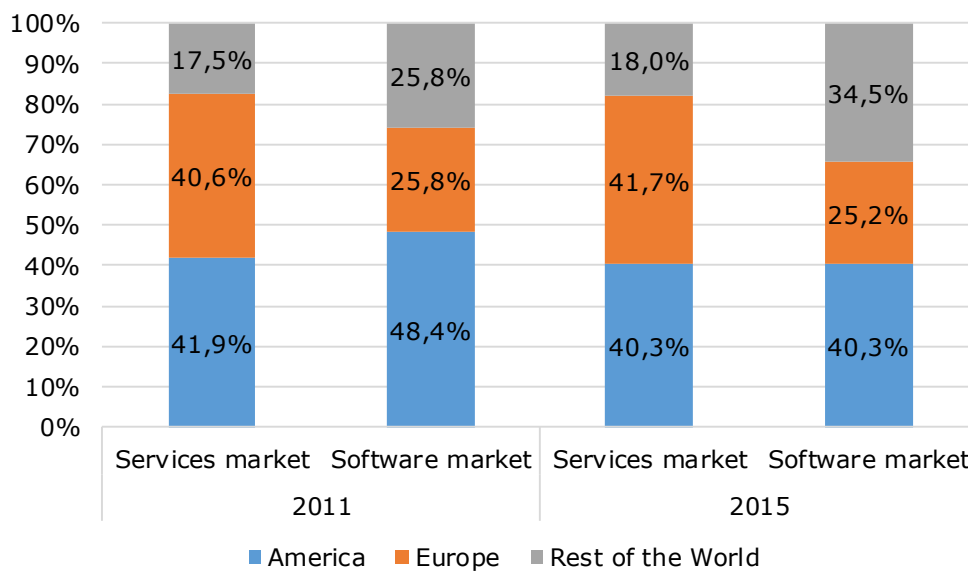
¹⁴ <http://lr-coordination.eu/cef>

¹⁵ www.lr-coordination.eu

Figure 8: Worldwide LT market forecast (billion euro)

Source: (LT-Innovate, 2016b)

Regarding the software sales, America accounted for 48.4 % of total revenues in 2011, while Europe represented 25.8 %. The distribution remained similar in 2015, although America's market share was expected to decrease to 40.3 %. Both regions reached similar market share considering LT services (41.9 % America and 40.6 % Europe). In this case, it was expected that the European market grew at a higher pace, reaching 41.7 % of global market in 2015, while the American market decreased to 40.3 % as can be seen in Figure 9.

Figure 9: LT revenues by regions (percentage)

Source: (LT-Innovate, 2013)

The demand for LT services is expected to grow in Europe. Currently, the European market is competitive with respect to American suppliers because of its stronger academic base; however, business adoption is faster in America.

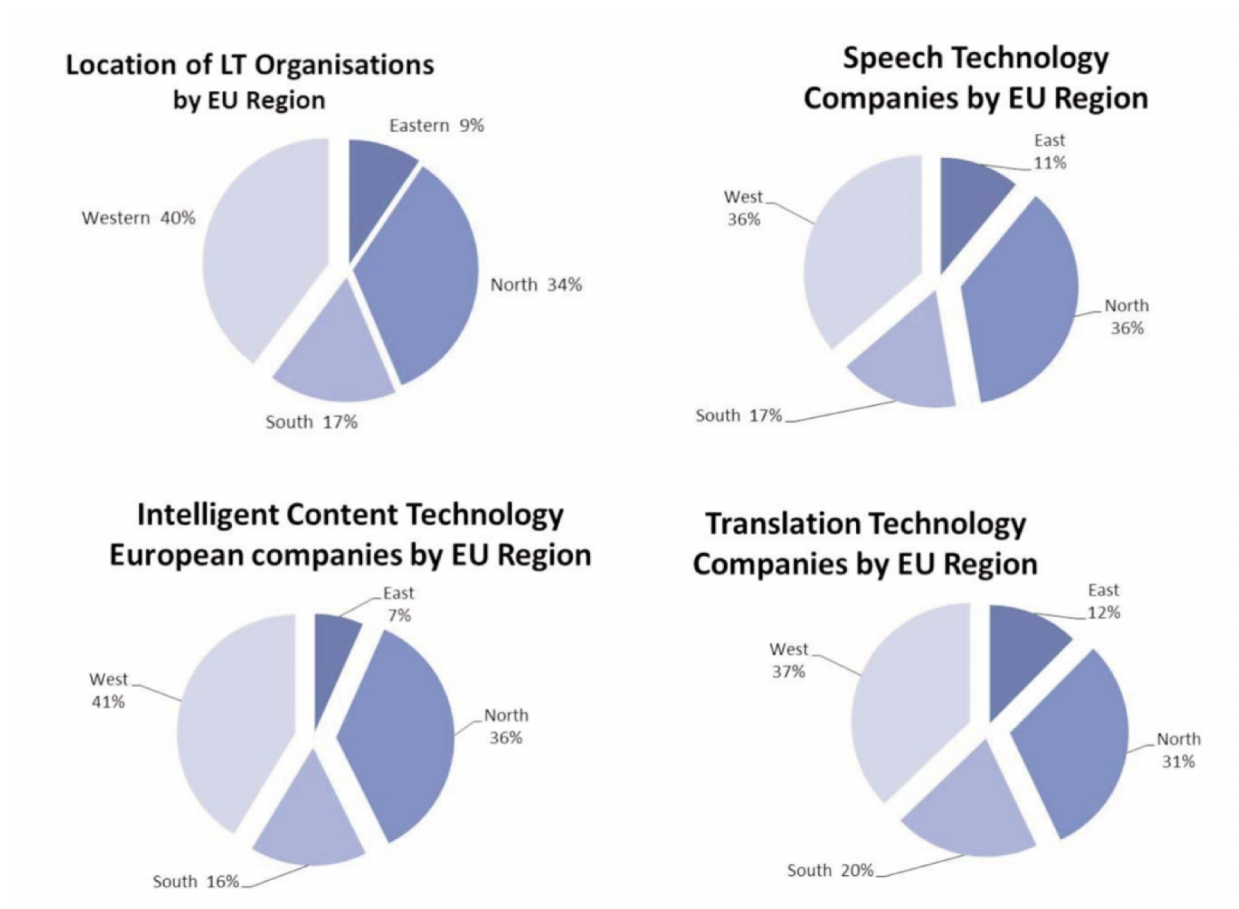
Table 4: Leading technology suppliers in the global market

Technology	Main providers
<i>Speech technology</i>	Microsoft – speech embedded in its software platforms
	Google – speech-enabled search
	Nuance – Enterprise and packaged solutions provider in the US and Europe
	iFlytek – 5,000 partners, owns 70 % of the Chinese-language ASR market
<i>Translation technology</i>	Google – free online translation and API for developers
	Microsoft – free online translation and API for developers
	Youdao – free translation in Chinese search engine
<i>Intelligent Content technology</i>	Endeca/Oracle
	FAST/Microsoft
	Google Search Appliance
	Lucene/SOLR
	Vivisimo/IBM
	Autonomy/HP

Source: Compiled by authors based on information from (LT-Innovate, 2013)

Regarding the European industry, the (Rehm et al., 2014) report estimates that there are around 500 companies in Europe either actively developing language technology or embedding its features in their products and services in an innovative way, i.e., the 2 % of the ICT industry.

The industry in Europe comprises of mostly small companies, as shown in Figure 11, concentrated in the western and northern regions of the EU, displayed in Figure 10, with a mix of long-established players and a significant number of new entrants. A quarter of the companies are micro-enterprises with fewer than 10 employees, while only 6 % had more than 200 employees; almost the entire industry is composed of SMEs. Over half the industry is comprised of companies that have been active for more than 10 years.

Figure 10: Location by EU region of LT industry, globally and by sectors

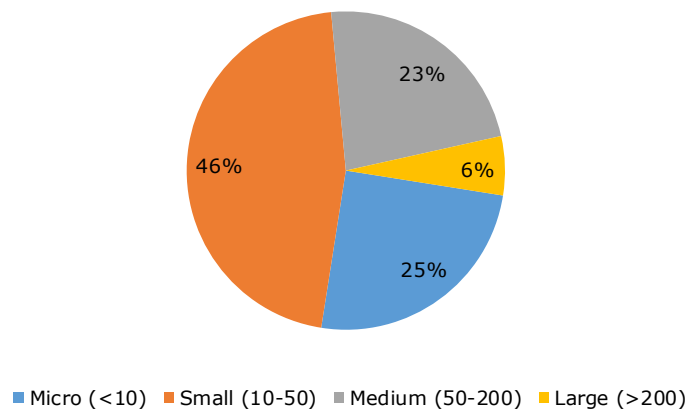
Source: Figures taken from (LT-Innovate, 2013).

NORTH: Denmark, Estonia, Finland, Ireland, Latvia, Lithuania, Sweden, United Kingdom.

SOUTH: Greece, Italy, Malta, Portugal, Slovenia, Spain.

EAST: Bulgaria, Cyprus, Czech Rep., Hungary, Poland, Romania, Slovakia.

WEST: Austria, Belgium, France, Germany, Luxembourg, Netherlands

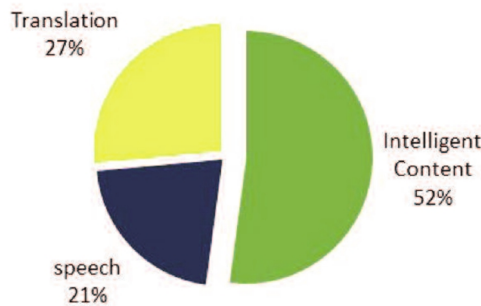
Figure 11: LT European industry by size (number of employees)

Source: (LT-Innovate, 2013)

As it can be seen in

Figure 12, most companies are focused on intelligent content, trying to leverage the potential of big data technologies.

Figure 12: European LT vendors by segment



Source: Taken from (LT-Innovate, 2013)

The trend of industry consolidation through acquisitions (described below) has deteriorated the European speech industry, reducing it to companies developing core technology and companies specializing in smaller languages (LT-Innovate, 2012a). The fact that so many companies fail to scale, even after years in business, is unusual in a technology industry, and indicative of the market context for LT in Europe: local/national companies with expertise in local languages serve local markets with services based on their own languages. This state of affairs is not likely to be sustainable, as cloud-based language-enabled services are launched on a large scale. At present, only a few number of European HLT companies – for example Expert System in Italy, Inbenta and Unbabel in Spain, Creative Virtual in the UK and Acrolinx in Germany (LT-Innovate, 2016b) – can be considered as global companies capable of competing in an ecosystem where ‘access to technology, rather than narrow linguistic expertise, is the driving factor’ (LT-Innovate, 2013 page 25).

Many European LT companies have gained access to markets by being acquired by larger companies, mainly American, which are seeking to complete their portfolio of products and services by acquiring small innovative start-ups, some of them from Europe (DG Translation, 2009). There are several examples of small and medium-sized European technology companies that have been picked off by large US players: the Dutch semantic search company Q-go is now part of Oracle; UK intelligence analytics firm i2 is now part of IBM; Loquendo, the speech company spawned by Telecom Italia, is the most recent European acquisition by Nuance; Spain’s NeoMetrics Analytics is now part of Accenture; the speech search engine of UK-based Aurix is now part of leading contact centre supplier Avaya and another call centre company Syntellect acquired Fluency Voice Technologies; Amazon acquired the UK-based natural language processing start-up, Evi Technologies, and the Polish firm, Ivona; the most successful VoIP service, Skype, was acquired by Microsoft; and three of the most relevant IT companies, Google, Apple and Microsoft, have recently acquired three UK-based LT firms, Deep Mind, VocalIQ and SwiftKey, respectively. And of course, Europe’s most successful intelligent search company, Autonomy, is now part of HP’s natural-language offerings. However, not all acquisitions are going to the US: Dassault Systèmes acquired the French intelligent search company Exalead; Experian, the global information services group based in Dublin, bought the UK speech-verification company 192business; OnMobile (spinoff of the Indian IT services giant Infosys) acquired the French speech company Telisma; and Wolters Kluwer acquired the US special-domain semantics company Health Language Inc. SDL in the UK acquired Language Weaver, the main statistical MT engine in the market developed for

enterprise (or government) markets rather than as an online platform. The French Bertin Technologies acquired Vecsys and AMI software, and the Italian Al maviva acquired Pervoice (LT-Innovate, 2016b)

This trend contributes to weaken the position of European industry on the global market, hindering the possibility of creating larger European companies that can compete with American leading firms. As a result, today the most visible innovations in translation technology, and business models for translation, are largely driven from outside Europe.

This consolidation in the European industry is healthy, and helps LT move up into mainstream applications and markets. It does not, however, promote the evolution of a strong and self-sustaining LT industry across Europe. This is evidenced by the patchy language coverage of solutions in the speech and content markets which is a key constraining factor in Europe's share of those segments of the market.

3.2 Socio-economic implications of multilingualism

There is no doubt that the language diversity in Europe is a source of cultural richness for the European society. It has allowed us to become what we are today, as it has made possible the adaptation of all ecosystems to our continent (Skutnabb-Kangas, T., 2002). Language is not only a way of communication but also a way to express concepts and ideas that may not exist in other cultures. If a language dies, concepts and ideas can also die. Europe has been able to create and spread all over the world some of the main concepts and ideas that have contributed to the evolution of mankind. And it has been possible thanks to its language diversity.

However, it also creates serious challenges both to the European economy and European society, seriously limiting aspects such as cross-border commerce and business, workers' mobility, provisioning of public services at the European level and citizens' participation in the political process. It is expected that HLT can help preserve and foster current European cultural diversity while providing the tools to overcome these problems.

Multilingualism affects many aspects of our daily lives related to cross-border economic and social relationships; therefore, it is quite complex to make an accurate estimation of the comprehensive effect of multilingualism in the European economy and society. In order to keep this study manageable, an analysis focusing on those aspects where multilingualism is likely to have a substantial impact and HLT are likely to provide a feasible and effective solution to tackle the linguistic barriers has been made. The analysis focuses on qualitative aspects because while in some areas such as e-commerce, its impact is more direct and may be to some extent quantified, in other areas, such as e-government, health, citizen engagement and the ability to close business deals, the effects are more diffuse and harder to estimate.

Although there is a strong relationship between economic and social effects, the main impacts have been classified into two categories.

The social aspects analysed are the following:

- How multilingualism affects European countries and, particularly, who are the citizens being left behind by using linguistic policies focusing on major languages. This analysis will be the basis to estimate the social and economic cost of non-multilingualism in Europe.
- The effect of language barriers on migration, particularly within the EU. To do so, the number of EU citizens living in the country and not speaking one of the official languages of the country is quantified. The effect of language barriers on internal mobility within the EU is also estimated.
- The effect of a monolingual regime where basic public services are provided by the governments only in the official language(s) of the countries (national and regional). The

analysis is focused on e-government services, health and emergency services that are likely to be the most relevant in the daily life of citizens and particularly affected by language barriers.

- Eventually, how HLT can help create a truly multilingual Europe that can further improve European construction by fostering citizens' engagement and participation is discussed.

The economic aspects analysed are the following:

- The most relevant effect of a non-multilingual regime in the DSM is lowering cross-border e-commerce. An overview of the trends and main barriers regarding cross-border e-commerce to further assess whether there is a true integrated DSM or a fragmented DSM is provided. The effect of linguistic barriers on e-commerce and what would happen by overcoming language barriers is estimated. Eventually, the main findings of previous research about the cost for Europe of not having an integrated DSM are summarised.
- A specific analysis of the impact of not using HLT for European SMEs is made.
- The problems of making business with countries using different languages and how international trading could benefit from using HLT.

3.2.1 Socio-demographic consequences of a non-multilingual regime in a multilingual Europe

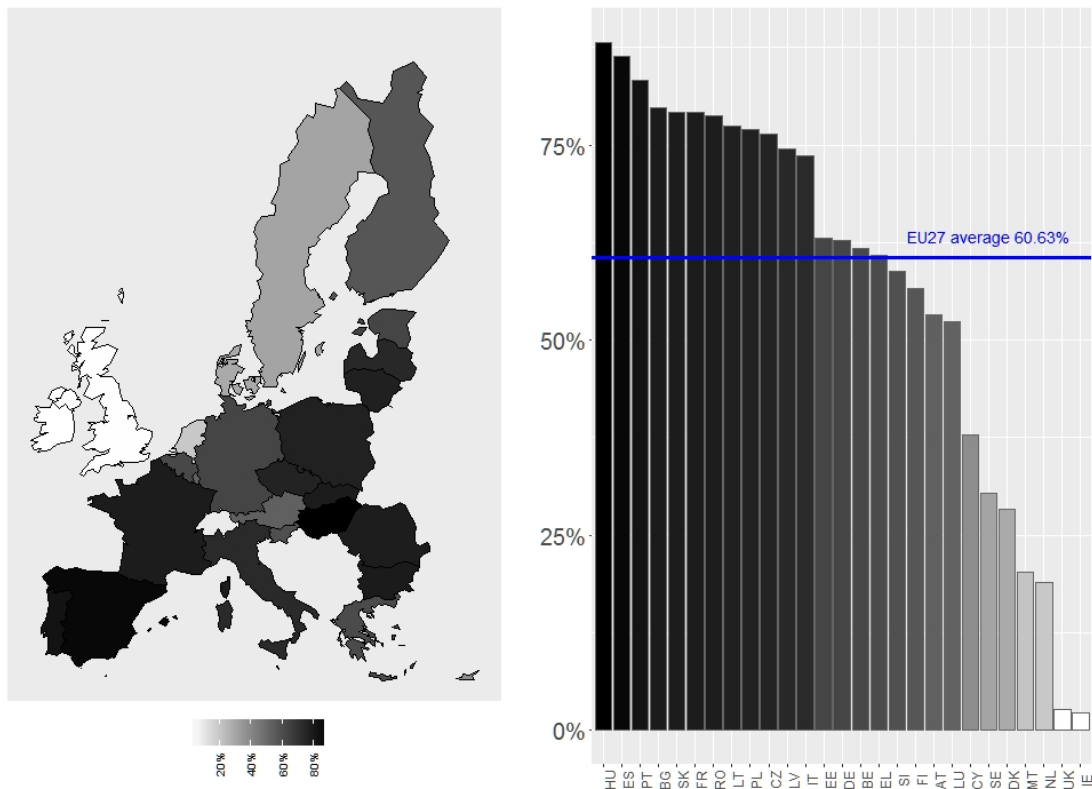
To properly assess the social and economic effects of not having enough HLT support in Europe to provide an efficient and straightforward way for citizens and firms to tackle the complexity of a multilingual environment, regardless of their ability to speak more than one language, it is important to know the population affected by the problem and whether or not linguistic barriers depend on the socio-demographic characteristics of individuals.

To do so, an analysis to estimate the population by country that is left behind in a monolingual and a multilingual scenario has been made. In the monolingual scenario we consider those EU citizens who speak English. In the multilingual scenario we consider those EU citizens who speak, at least, one of the six most spoken European languages (English, German, French, Italian, Polish or Spanish)¹⁶.

The analysis is performed by country, both for the whole population and by different socio-demographic characteristics (age, income level and age leaving education) to capture an idea of the language divide.

The results of the monolingual scenario show that at least, 60 % of the European population is left behind when using only English, and with high disparities among the countries, as can be seen in Figure 13. In countries such as Hungary, Spain, Portugal and Bulgaria less than 20 % of the population is able to speak English compared to 80 % in the Netherlands.

¹⁶ To make the analysis microdata from the Eurobarometer 77.1 (European Commission, 2014a) have been used. The survey was performed in February-March 2012 to 26.751 citizens in the 27 countries of the European Union in 2012. All respondents were residents in the respective country, nationals and non-nationals but EU-citizens, and aged 15 and over. The survey also includes demographic and other background information such as age, gender, age when stopped full-time education, level in society, and type and size of locality. The Eurobarometer 77.1 includes a special topic on Multilingualism with information of the mother tongue and the level of knowledge of foreign languages (in of a scale of three: *very good*, *good*, and *basic*). We consider that a citizen speaks a language whether the language is its mother tongue or whether the citizen has, at least, a good knowledge of the language as a foreign language.

Figure 13: Percentage of population not speaking English by country

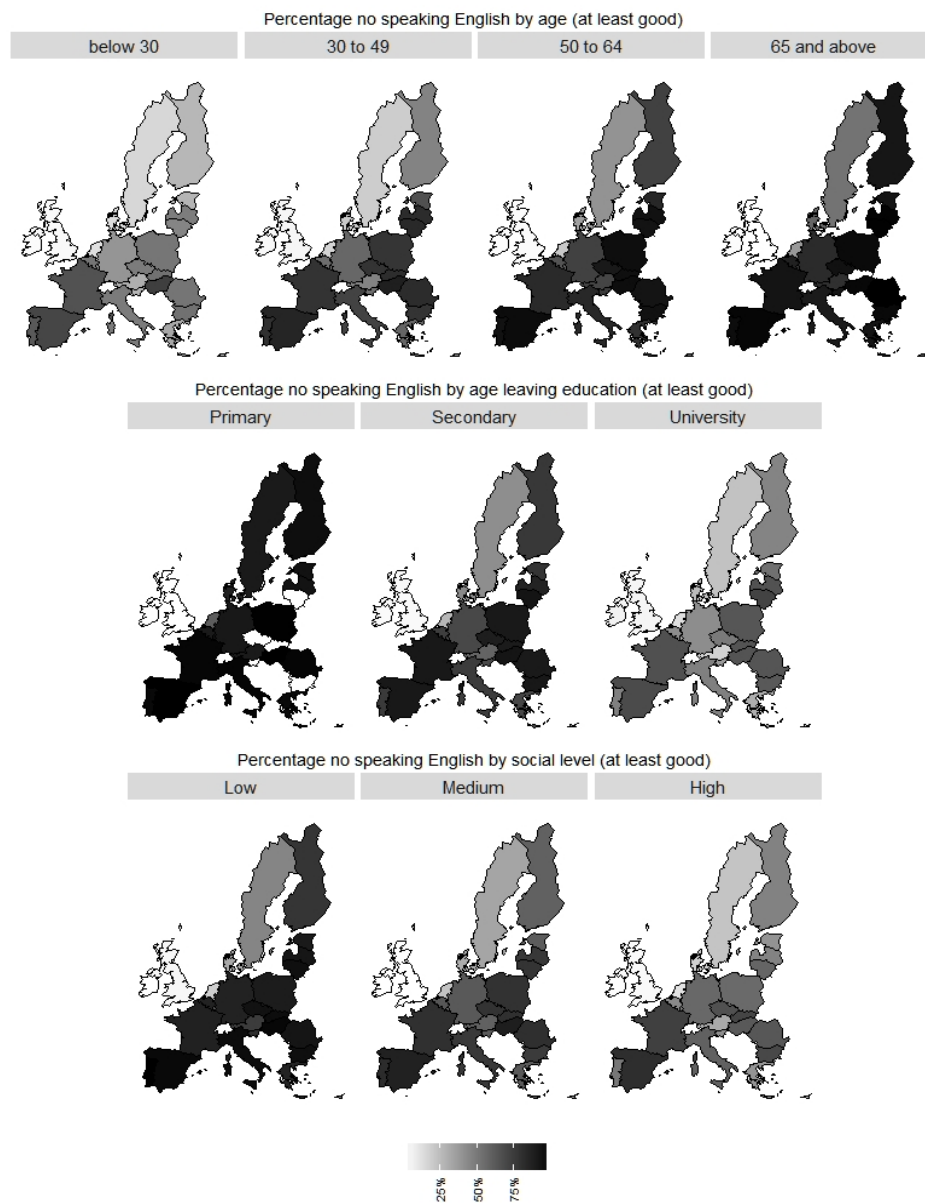
Source: Compiled by the authors based on the Eurobarometer 77.1 (European Commission, 2014a)

The analysis also shows that the linguistic gap affects older, less educated, and low income populations more as can be clearly seen in Figure 14. The age gap between populations younger than 30 and populations older than 66 is 34 percentage points. It is important to remark that even for younger populations, when considering only English, the language barrier at the European level is higher than 40 %.

The language gap also depends heavily on education level¹⁷. While over 80 % of less educated populations have a low level of English, the figure goes down to about 40 % when considering populations that left education after the age of 19 (they are likely to have some university education). The education gap average is 39 percentage points. Very similar results can be found when considering the social level,¹⁸ with an average gap of 42 percentage points between low and high social levels.

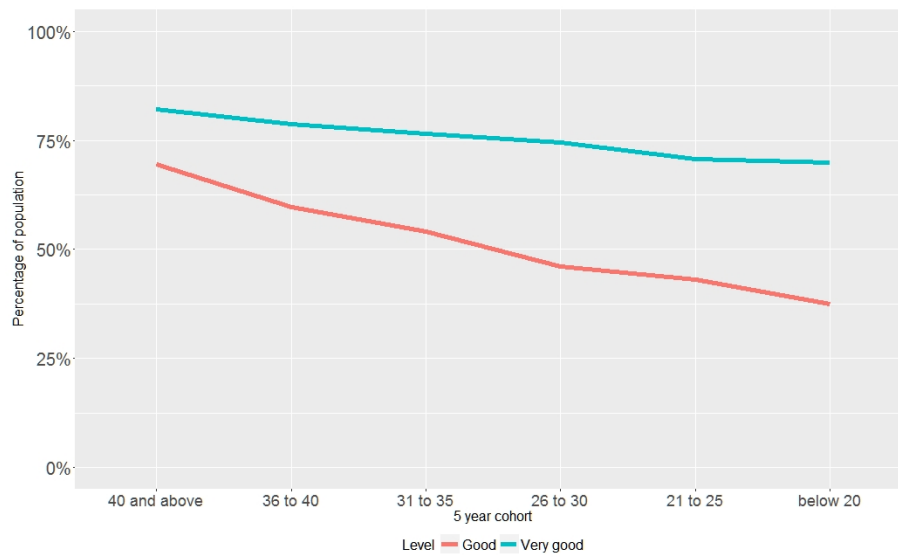
¹⁷ We consider that an individual has Primary education if leaving education before 16, Secondary education if leaving education between 16 and 19, and University education if leaving education after 19.

¹⁸ The survey considers 10 social levels ranging from 1 (the lowest) to 10 (the highest). These levels are grouped into low (1 to 3), medium (4 to 7), and high (8 to 10).

Figure 14: Percentage of population not speaking English by different socio-demographic factors

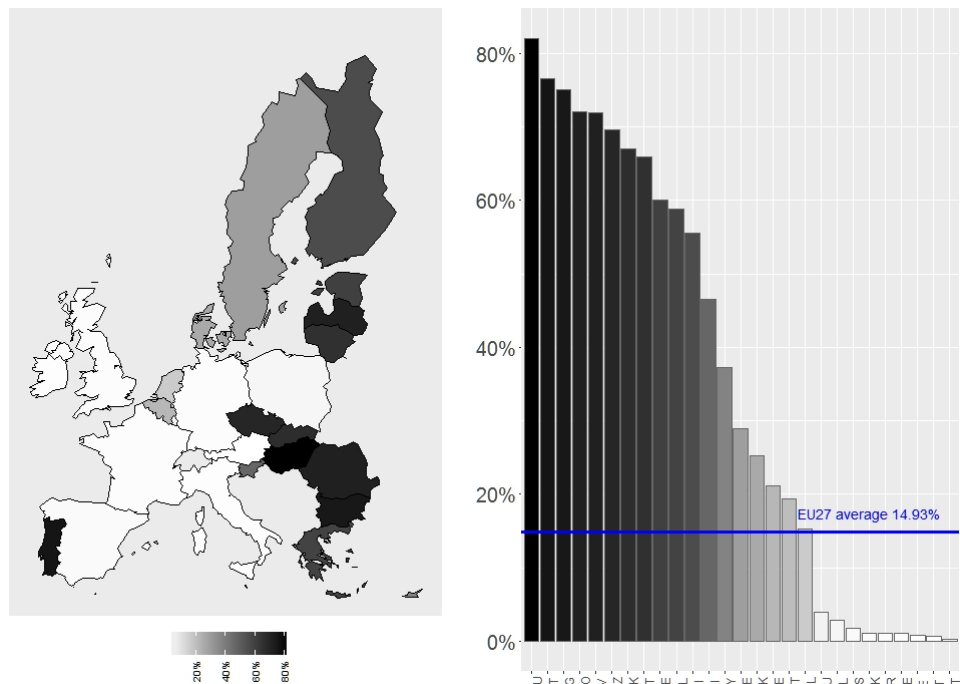
Source: Compiled by the authors based on the Eurobarometer 77.1 (European Commission, 2014a)

Moreover we should not expect the situation to improve substantially in the coming years. By analysing the percentage of populations not speaking English by 5 year cohorts, we can assess how fast younger populations will improve their language skills. There is still 35 % of the younger population that is not speaking English, at least reasonably well, and 70 % that is not speaking very well. The situation is improving slowly with 22 percentage points having a good level and only 9 percentage points having a very good level in the last 20 years. Moreover, for those speaking English very well the improvement shows symptoms of stagnation, as seen in Figure 15. These results are in line with the findings of Gazzola (2016), suggesting that the level of language skills of European students is deceptive and need to be substantially improved.

Figure 15: Evolution of percentage population in 5 year cohorts not speaking English by level

Source: Compiled by the authors based in the Eurobarometer 77.1 (European Commission, 2014a)

Moving towards a multilingual scenario of six languages substantially improves the situation as expected. However, 15 % of the European population is still left behind and there are 11 European countries (Finland, Greece, Estonia, Lithuania, Slovakia, Czech Republic, Latvia, Romania, Bulgaria, Portugal and Hungary) with more than half of the population unable to properly communicate using any of the languages listed in Figure 16.

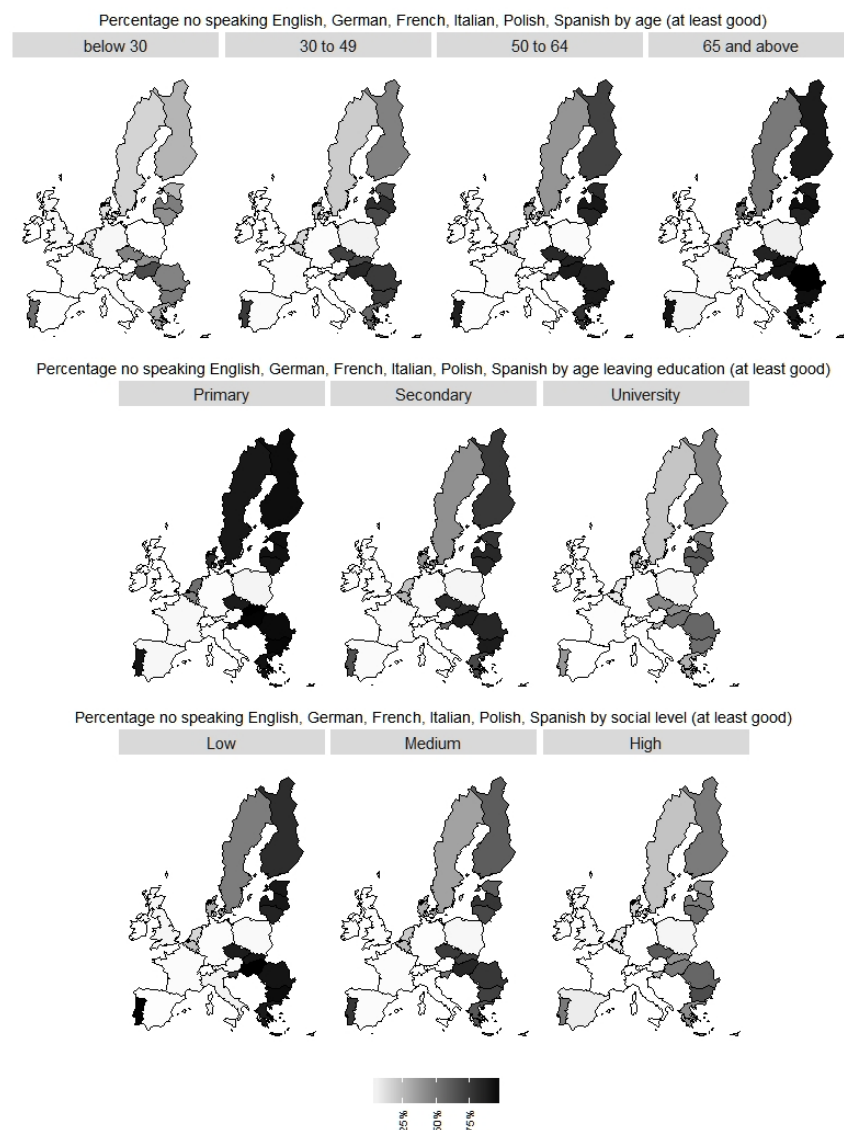
Figure 16: Percentage of population not speaking English, German, French, Italian, Polish or Spanish (at least good) by country

Source: Compiled by the authors based in the Eurobarometer 77.1 (European Commission, 2014a)

These results are in line with previous research. Gazzola (2016, p.7) makes a similar analysis and also summarises the results of other works to find that the “intermediate disenfranchisement rate”, a similar indicator to the one we have calculated, is between 62.6 and 65 (our estimate is 60.63) for a monolingual regime and between 14 and 16.4 for a six languages regime (our estimate is 14.93).

In a similar way to the monolingual scenario, the results show that the linguistic gap affects older, less educated, and low income population much more as can be seen in Figure 17. The age gap between populations younger than 30 (11 %) and population older than 66 (19 %) is eight percentage points. It is important to remark that even for younger populations, when considering the six most spoken European languages, the language barrier for these languages at the European level is 11 %.

Figure 17: Percentage of population not speaking English, German, French, Italian, Polish or Spanish by different socio-demographic factors



Source: Compiled by the authors based in the Eurobarometer 77.1 (European Commission, 2014a)

The language gap also depends heavily on education level. While more than 19 % of the less educated population have a low level of one of the six most spoken European languages, the figures go down to 10 % when considering populations that left education after the age of 19 (an average gap of 9 percentage points). When considering the social level, we get an average gap of 16 percentage points between low (28 %) and high (12 %) social levels.

The different socio-demographic factors are likely to be strongly intertwined (for instance higher level society populations are more likely to have university studies and live in urban areas) and therefore it is particularly interesting to analyse all the variables together to disentangle the different effects and to establish a more accurate relationship between the socio-demographic factors and the language skills. The detailed results of the analysis are included in Annex 7.3. The results suggest that education is the characteristic more influential in relation to speaking a foreign language. On average, leaving education after age 19 increases the chances of speaking English as a foreign language about 19 times and the chances of speaking one of the six most spoken European languages about nine times, compared to leaving education before age 16. The age is also relevant and younger people (below 30) have, on average, five more chances of speaking English as a foreign language and four more chances of speaking one of the six most spoken European languages compared to older people. The effect of the social status, although important (higher social status increases the chances to speak English three times more compared to lower social status), is lower than education level. The influence of the size of the place of residence is significant, but much smaller, and the effect of gender is negligible.

In summary, our findings suggest that if English becomes the “lingua franca” of the DSM, more than 60 % of the European population will be left behind and with high disparities between countries. Moving towards a scenario where the six most spoken European languages become the means of communication in the DSM, more than 14 % of the population will be left out of the DSM. That is more than 70 million European citizens. There will be 11 EU countries with more than 50 % of its citizens unable to access the advantages of the DSM. To compound the problem, the citizens who will be more likely to be disadvantaged are older and less educated, creating an unfair linguistic divide in the DSM.

3.2.2 Consequences of a non-multilingual regime within the countries for migrants and cross-border mobility

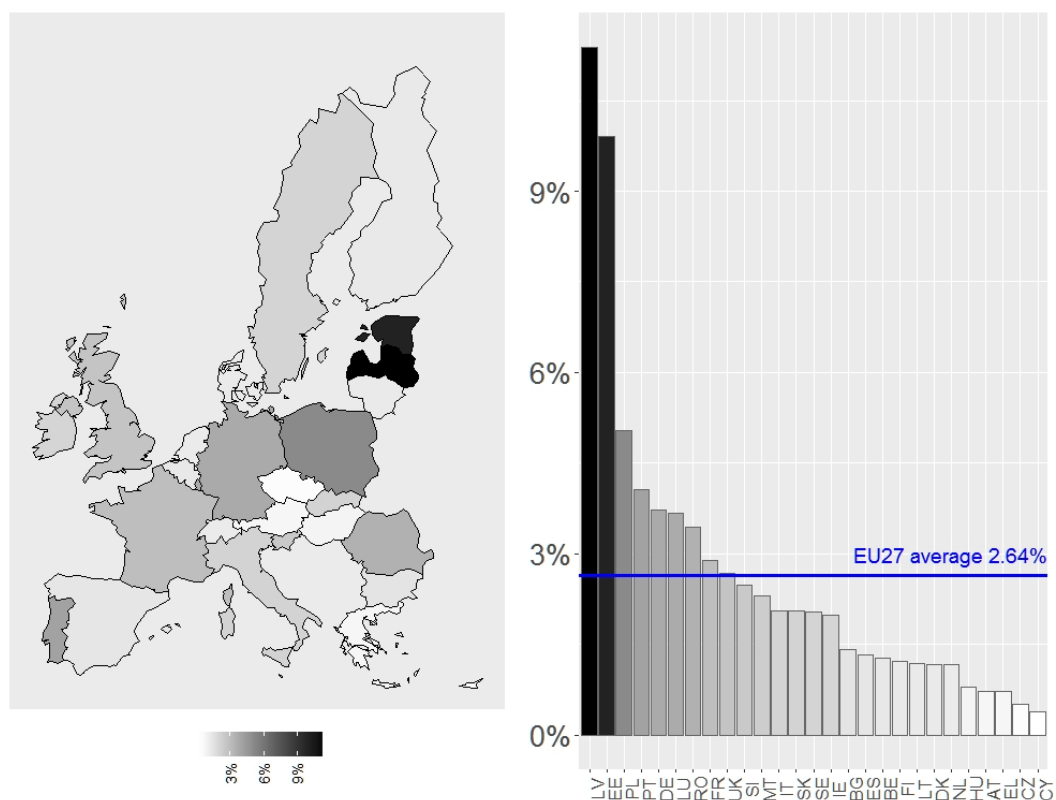
According to the United Nations’ International Migration Report (UN, 2016), nearly two thirds of all international migrants (76 million) live in Europe. In the new global environment, language may become a difficult challenge for integrating migrants because living in a country without properly speaking the official language imposes a burden on migrants, who are unable to find a job or successfully access basic public services such as e-government, health or emergency services.

For these citizens, leading a normal life is likely to be challenging. To compound the problem, migrant languages are the least recognised in the policies of European countries, challenging the mobility and true integration of Europe (Extra, Guus & Kutlay, Yağmur, 2012). HLT can substantially improve the integration process by facilitating the communication of newcomers (META technology council, 2012).

Regarding internal migration (EU citizens living in another EU country), there are more than 13 million European citizens (2.64 %) living in an EU country who are unable to speak the official language of the country¹⁹, at least reasonably well, as seen in Figure 18. It is likely that most of these citizens are (or have been) internal migrants.

¹⁹ The list of official languages by country is included in Annex 7.2.

Figure 18: Percentage of population not speaking any of the official language(s) of the country by country



Source: Compiled by the authors based in the Eurobarometer 77.1 (European Commission, 2014a)

HLT will not only facilitate the integration of migrants but will also foster a more integrated European Union because language differences are expected to be a barrier for EU workers moving to a different Member State. To check that fact we have estimated the effect of language barriers on the mobility of EU citizens within the EU to a different Member State to live and work by using the Eurobarometer 72.5 dataset (European Commission, 2009) that includes a specific topic about geographical and labour market mobility²⁰. We have considered a mobile worker EU citizen, a citizen that is currently working in a Member State different to its citizenship or that has done so in the past. The percentage of EU citizens who have ever moved to a different Member State to live and work within the EU is 5.8 %. We have found in our analysis that language differences are likely to be one of the main barriers hindering workers mobility within Europe and challenging the creation of a truly single market. On average, having low language barriers between two countries increases the population that decides to move between those countries by 118 % (more than three times) as shown in the analysis of Annex 7.5.1.

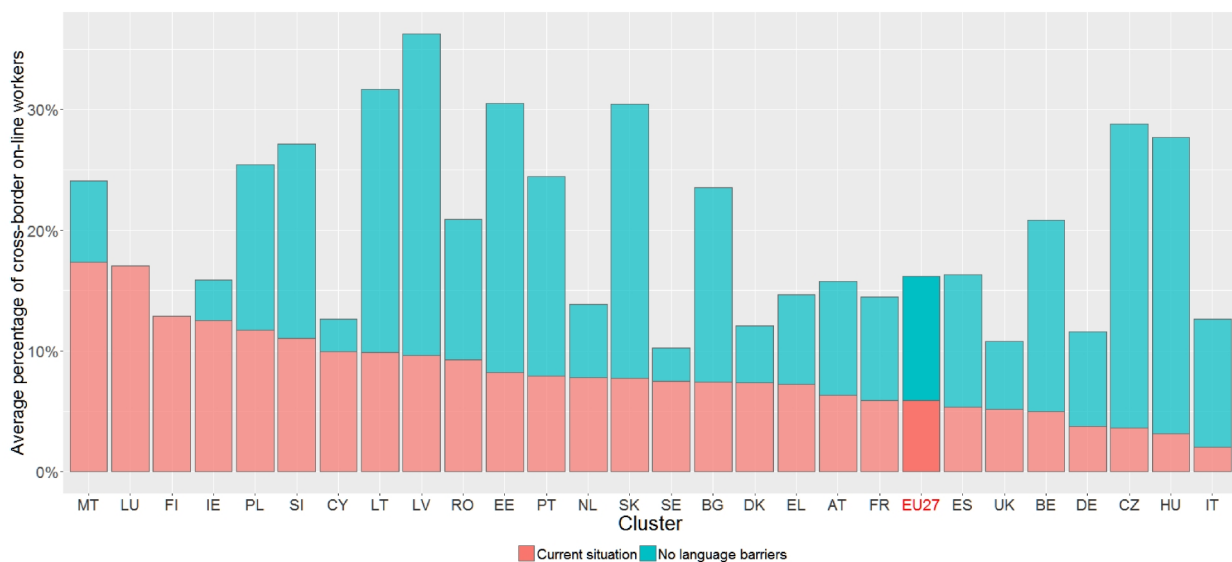
We have also made a simulation to predict the effect of having no language barriers. In a scenario where there are low language barriers between countries, and based in our model, this figure would increase almost three-fold (2.8) up to 16.1 % as can be seen in Figure 19²¹.

²⁰ A detailed description of the analysis is included in Annex 7.5.1.

²¹ These results should be taken carefully because the lack of language barriers variable in our model may be capturing other cultural barriers between the countries that are not adequately included in our model. However we

In fact, in 2014 only 1.8 million Europeans (0.26 %) migrated to a different Member State (European Commission, 2016j), compared to 5 million of people (1.63 %) that moved to a different State within the USA (USCB, 2015). The internal migration rate in the USA is therefore 6.4 time higher compared to the EU. Based on our analysis, lowering language barriers could halve the working mobility rate difference between the EU and the USA.

Figure 19: Percentage of population that have lived and worked in another EU country (current situation compared to having no language barriers)



Source: Compiled by the authors based on (European Commission, 2009, 2014a; EUROSTAT, 2016c)

3.2.3 The impact of non-multilingual Europe on providing public services

One of the reasons that could partially explain why language barriers are so important for EU mobility is the lack of public services in the destination country provided in the language of the EU migrant. In this chapter we analyse the fairness in the access of EU citizens to three public services: health, emergency and e-government services

3.2.3.1 Health and emergency services

Health is a public service where multilingualism plays a relevant role. Both patients who do not speak the official language of the Member State where they live (for instance migrants) and patients who travel to other Member States require language support when they access medical care – information in their own language, translation of documents, interpretation to receive safe healthcare, etc. (European Commission, 2015). At European level, the Directive 2011/24/EU encompasses the patients' rights in cross-border healthcare, although it does not explicitly state the patient's right to communicate in a language they understand when seeking medical care, taking it for granted as a prerequisite to ensure the rest of the rights included (European Commission, 2015). This lack of legal definition about how to

think that provides evidence that language barriers are very likely to have a profound effect on the low working mobility within the EU.

bring down language barriers in cross-border healthcare leads to an unbalanced provision of translation and interpreting services in national healthcare systems which, in turn, can lead to economic inefficiencies and high social costs such as the loss of human lives. In many cases, the provision of language services is still not considered a key element of the healthcare system, but a mere economic problem (European Commission, 2015) that has conveniently not been addressed in recent years due to budget cuts derived from the economic crisis.

Among the impacts of language barriers on the provision of health services it is worth to point out the following (Bowen, S., 2015):

- barriers on participation in health promotion and prevention activities;
- barriers to initial access for most health services;
- increased risks of misdiagnosis;
- poorer patient understanding and adherence to prescribed treatment;
- lower patient satisfaction;
- increased risks of experiencing adverse events;
- poorer management of chronic disease;
- less effective pain management.

One indicator that allows highlighting the cost of non-multilingual healthcare systems is the number of medical errors derived from language barriers. Although literature at European level has not been found, there are examples from other regions that illustrate this issue. A study carried out by the School of Public Health of the University of California analysed the incidence and consequences of language barriers in medical malpractice claims of an insurance carrier for USA States (Quan, K. & Lynch, J., 2010). They discovered that 2.5 % of claims registered between January 2005 and May 2009 had to do with language barriers (in this case, patients with limited English proficiency), with the result of two children and three adults dead and more than \$5 million paid for damages and legal fees.

Related to health and security, the multilingualism also has a relevant impact in the provision of emergency services. The single European emergency call number, 112, was introduced in 1991 (Council of the European Communities, 1991), and the Universal Service Directive, enacted in 2002 and amended in 2009 (European Parliament and the Council of the European Union, 2009), established the requirements for Member States concerning 112 services. However, the regulation of 112 services did not explicitly address the multilingual handling of the calls, limiting to establish that 'Member States should have already made the necessary organisational arrangements best suited to the national organisation of the emergency systems, in order to ensure that calls to this number are adequately answered and handled' (European Parliament and the Council of the European Union, 2009). According to the results of the sixth data-gathering round on the implementation of the European emergency number 112 (European Commission, 2013c), which gathers data from 2012²², each Member State decides in which languages the 112 calls can be handled and the procedures to transfer the calls between PSAPs (Public Safety Answering Points) in order to provide a better service handling the call in the appropriate language. Although HLTs can help to conveniently deal with 112 calls, for instance reducing the time to identify the language used in the call, only a few countries reported the use of some kind of technologic support when handling emergency calls (European Commission, 2013c). One of the consequences of the disparity of approaches to handle foreign language calls and the scarce use of human language technologies is that 12 % of people who made an emergence call while visiting another EU country experienced language-related problems (Eurobarometer, 2010).

²² There have been two more recent data-gathering rounds, although they did not address the analysis of the calls in foreign languages

3.2.3.2 e-government

In a digital single market with free movement of citizens and goods, public administrations should provide efficient and cross-border citizen-centric services. However, the European Union's internal market is fragmented, also regarding e-government services. Information, advice, problem-solving mechanisms and procedures are dispersed, not user-friendly and lack interoperability (European Commission, 2016a).

Nowadays there are still many barriers to the mobility of workers and citizens and to the creation of businesses in other Member States. We have already analysed that language is likely to be one of these barriers. Therefore, digital public services need to address this challenge to truly achieve the objective of being effective and accessible for all citizens, and contribute to the creation of the Digital Single Market in a significant manner.

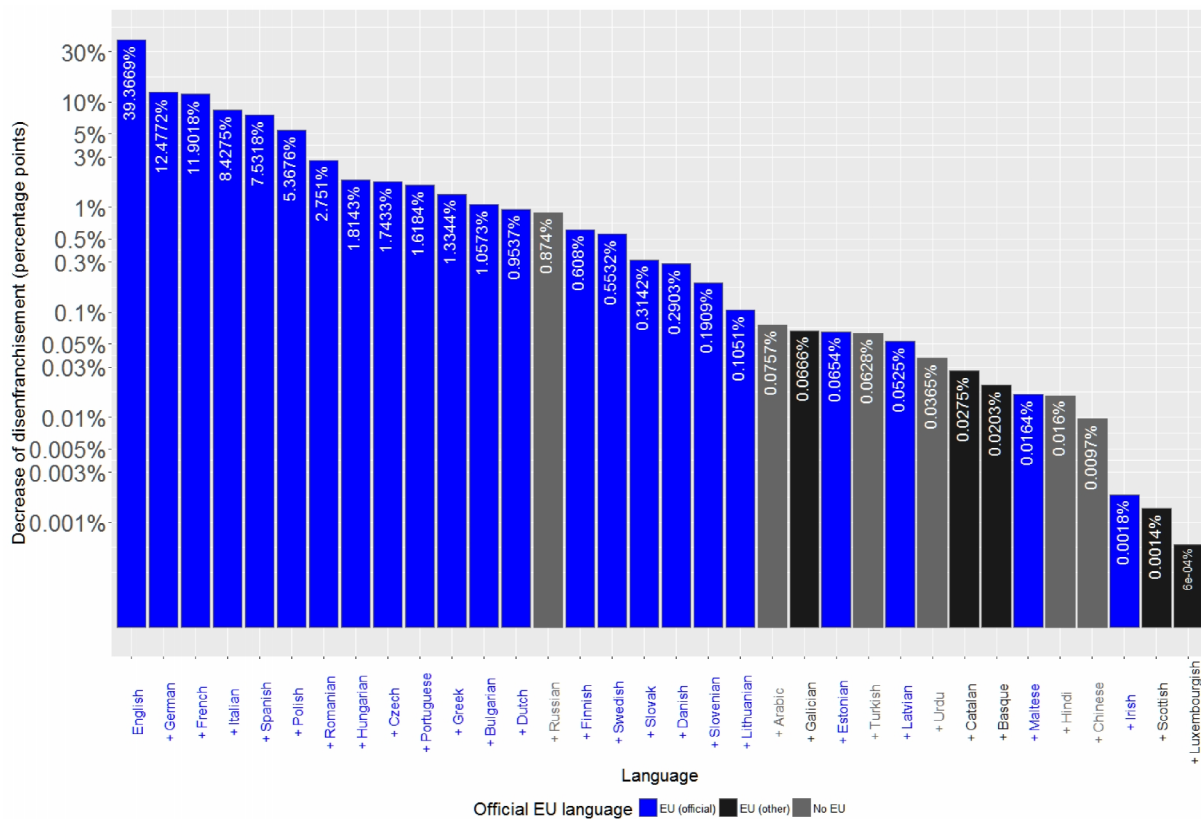
A study from 2013 carried out in the frame of the ISA Work Programme (European Commission, 2013b) found out that 34 % of public administration portals of Member States analysed only provided information in the official languages of the country and 66 % offered information in other languages, with English being the most common (61 %).

Multilingualism has two faces when it comes to e-government services. On the one hand, it is essential to guarantee that all citizens can access public services in their own language. In a context where a citizen has the right to live, study or work in any country of the Union, should not a citizen be able to understand the laws or access the public services of another country? In the digital era, user-centric services should also mean multilingual services.

On the other hand, multilingualism affects the semantic interoperability of ICT services and tools. Seamless interoperability between government and users – citizens, business, and other public administrations – both at the national and the EU levels require multilingual ontologies, which are key for semantic interoperability.

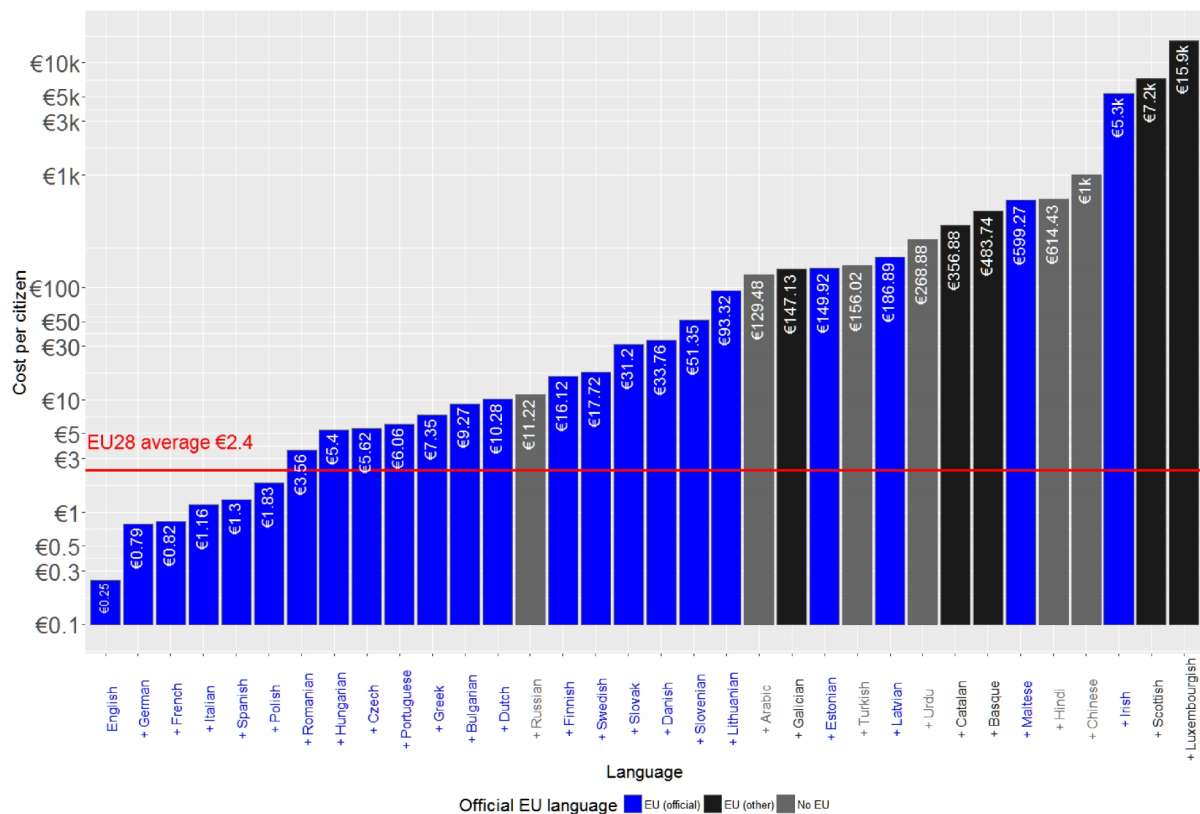
Machine translation tools appear as the only cost-effective way of making public services all around the EU available in all official languages and other spread languages among European citizens, such as Russian. This fact can easily be checked by analysing the cost of the translation services of the European Union. In order to fulfil the obligation of respecting the linguistic diversity of European citizens, the EU spends an estimated 1.1 billion euro per year (2012), accounting for 1 % of the annual general budget of the EU (Gazzola, M. & Grin, F., 2013). Gazzola, M. (2014) suggests that the current full multilingual policy of the EU, is not only inclusive, but it is also the most effective language policy because the yearly expenditure per citizen, of at least 15 years old, is about 2.7 euro. However, Fidrmuc (2011) claims that the average cost should be calculated depending on the number of citizens disenfranchised by excluding a particular language, resulting in a cost per disenfranchised person as high as 800 euro for Maltese. We have estimated the percentage of population that are no longer disenfranchised when adding a language²³ and the cost of adding a new language per person no longer disenfranchised and the results are shown in Figure 20 and Figure 21 respectively.

²³ We consider the population who speak the language at a level at least good. When adding a new language, we exclude the population that speaks one of the languages already considered at a level at least good.

Figure 20: Disenfranchisement rate when introducing a new language (logarithm scale)

Source: Compiled by the authors based in the Eurobarometer 77.1 (European Commission, 2014a)

It is true that the average yearly cost per person is close to 2.7 euro (our estimation is 2.4 euro). However, it is also true that the cost of including some smaller languages is as high as 5.300 euro per disenfranchised person for Irish or up to 15.000 euro per person for Luxembourgish. Moreover, it would be more effective to include non-official languages such as Russian, Turkish or Arab than some official languages. Fairness and effectiveness seem to be conflicting terms and HLT seems the only feasible solution to solve the dilemma.

Figure 21: Yearly cost per disenfranchised person per language (logarithm scale)

Source: Compiled by the authors based in the Eurobarometer 77.1 (European Commission, 2014a)

National and regional public documents from the different EU Member States are not usually available to other European citizens in their own languages. Take a Portuguese entrepreneur interested in the commercial law of Latvia or Slovenia. It is very unlikely that those documents will be available in Portuguese. Manually translating all the government documents of the 28 EU countries to all the languages of the EU is a utopia. This is particularly relevant in the digital age where physical barriers are overcome but cultural and language barriers remain. The only feasible solution is to use machine translation technologies that allow translating contents cheap and easily. That is why HLT are a crucial technology for multilingual Europe in the digital age.

3.2.3.3 Big data, open data and RPSI

Language is a type of data and it represents one of the most relevant challenges for the development of big data and the exploitation of its economic advantages.

Language technologies are the necessary link between big data and knowledge and are a part of the data value chain. In the EU, text content is created in various languages, and therefore text-based data is multilingual. This data, the resource of any big data application, is not currently being exploited because existing tools lack, in the majority of cases, the capability of processing languages other than English. For example, extracting knowledge from users' created content in Europe will not be possible in an effective way if the language barrier is not overcome. Accessing this data in a seamless manner thanks to HL will be very useful for developing business intelligence applications, data analytics, optimizing security solutions, etc. (SIRIA, 2016).

In this context of making the most of existing data in Europe, the role of the public sector is particularly relevant. The Re-use of Public Sector Information (RPSI) has also important links to HLT, as a data

source that is still to be maximized. The potential benefits of RPSI are not being exploited at the European level due to language barriers. Aggregating data at the EU level is very difficult and users cannot access an important part of data from another country and, sometimes, even from the same country due to these barriers.

HLT are the means through which public administrations can achieve the objectives of the RPSI, namely enhancing transparency and accountability and taking advantage of its economic and social gains.

RPSI is, at the same time, a source of information that will allow for the development of HLT. Public sector information, when made public with open standards and machine-readable formats, is a key resource for the development of new language technologies and the creation of language models.

3.2.4 A truly multilingual Europe fostering the European construction, citizenship engagement and reinforcing a common identity

Achieving open government in Europe, where public services are not only more effective and efficient but inclusive, would not be possible in the current context of language fragmentation. Citizens and interest groups would only engage in a meaningful conversation with decision-makers if it is done in their native language; otherwise, an important part of the social intelligence is lost. Participation and collaboration through online tools and the extraction of knowledge from social media and online content in Europe require that information and data are fully provided in all official languages and that citizens can address national and European authorities and other stakeholders in their native language (European Commission, 2015g).

According to the Eurobarometer 84.1 (TNS, 2015), 51 % of European citizens are attached to the European Union, compared to 90 % of Europeans that are attached to their country or 88 % that feel attached to their region. Although the percentage of Europeans that feel attached to the EU has increased 6 percentage points compared to the previous 2014 barometer, recent events, such as Brexit, have put this issue at the center of the debate in Europe.

The existence of a sense of belonging to the European Union has been a concern since the beginning of the European integration process. Creating a common citizenship was understood since the origins of the Union as a key element to achieve prosperity and peace in the continent.

The construction of a European identity has been founded on a common historical heritage and shared values of respect for human dignity, liberty, democracy, equality, the rule of law and respect for human rights²⁴. However, Europe faces multiple challenges in achieving this goal, with the linguistic diversity of the Union being an important one. Although identity is a multidimensional and even dynamic concept (Graves-Brown, Jones, & Gamble, 2013), there is no doubt that being able to make oneself understood is critical to developing a feeling of belonging (Weiß & Schwietering, NA).

A common language allows for communication and fraternization of peoples, and, as a result, it helps others understand different people's history, cultural wealth or even circumstances.

The development of LT might foster mobility and information exchanges and could allow personal communication among citizens that do not share one language, and it might therefore have a positive impact in fostering mutual understanding between Europeans.

²⁴ Article 2 of the Treaty on European Union (TEU).

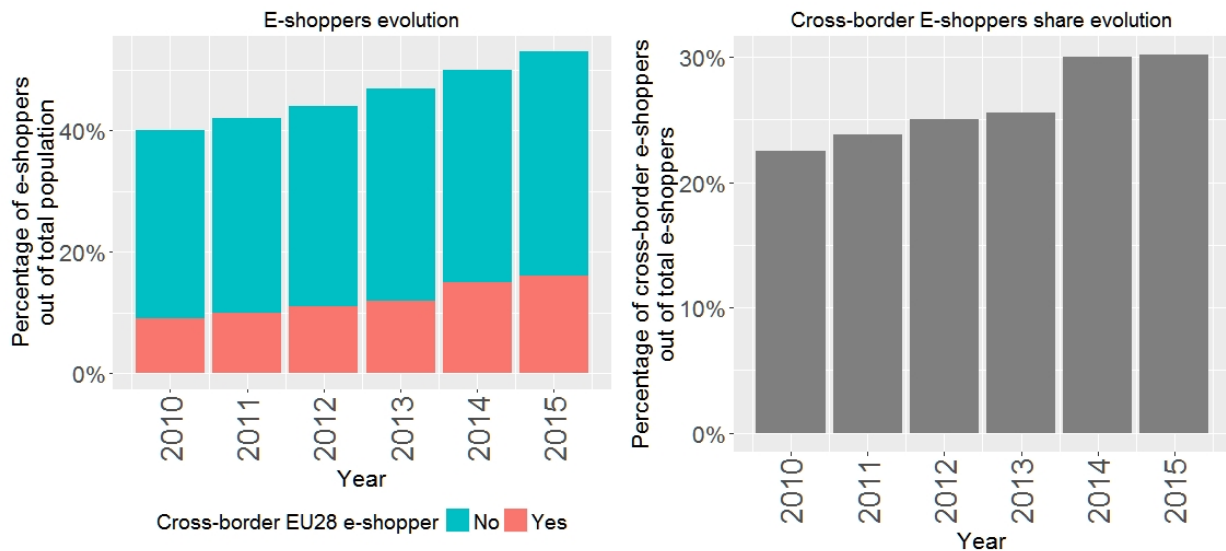
3.2.5 The effect of a non-multilingual Digital Single Market in cross-border e-commerce

One of the main objectives of the Digital Single Market (DSM) is boosting the e-commerce in the EU (European Commission, 2016k). However, language barriers are hindering the effective implementation of cross-border e-commerce and there lacks any mention in the Digital Single Market Strategy for Europe (EC, 2015) of policies intended to overcome them.

3.2.5.1 General trends in cross-border e-commerce

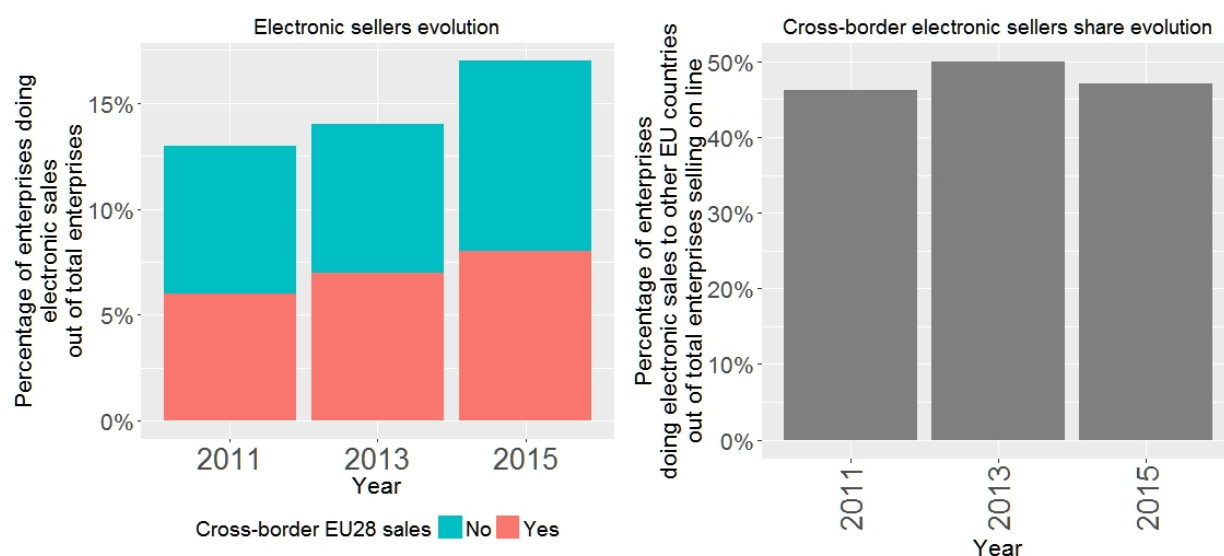
The percentage of European e-shoppers has steadily grown in the last years from 40 % in 2010 to 53 % in 2015, although the share of cross-border e-shoppers and total e-shoppers shows symptoms of stagnation as can be seen in Figure 22. In 2015, only 16 % of European citizens have purchased online from other EU countries, accounting for 30 % of total e-shoppers in that year (Eurostat, 2016).

Figure 22: Cross-border e-shoppers evolution



Source: Compiled by the authors based on (EUROSTAT, 2016a)

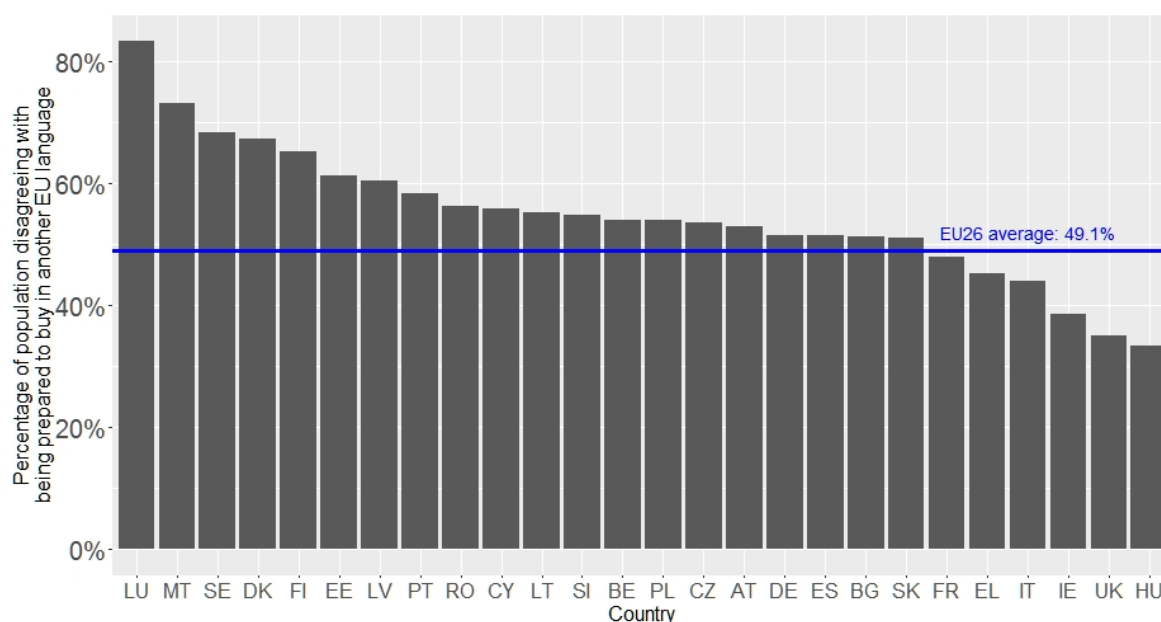
While considering the supply side, a similar pattern arises. The percentage of enterprises that are selling through electronic channels has slowly grown in the last few years from 13 % in 2011 to 17 % in 2015. However, the share of cross-border electronic sellers in total electronic sellers is stagnated at around 44 % as shown in Figure 23.

Figure 23: Cross-border e-sellers evolution²⁵

Source: Compiled by the authors based on (EUROSTAT, 2016b)

3.2.5.2 The perception of language barriers

Language differences are not the unique factor that explains this low penetration of cross-border e-commerce, although they are likely to play a relevant role. About 50 % of European consumers think that they are not prepared to buy in another EU language, with the figure ranging from 83 % in Luxembourg to 34 % in Hungary as shown in Figure 24.

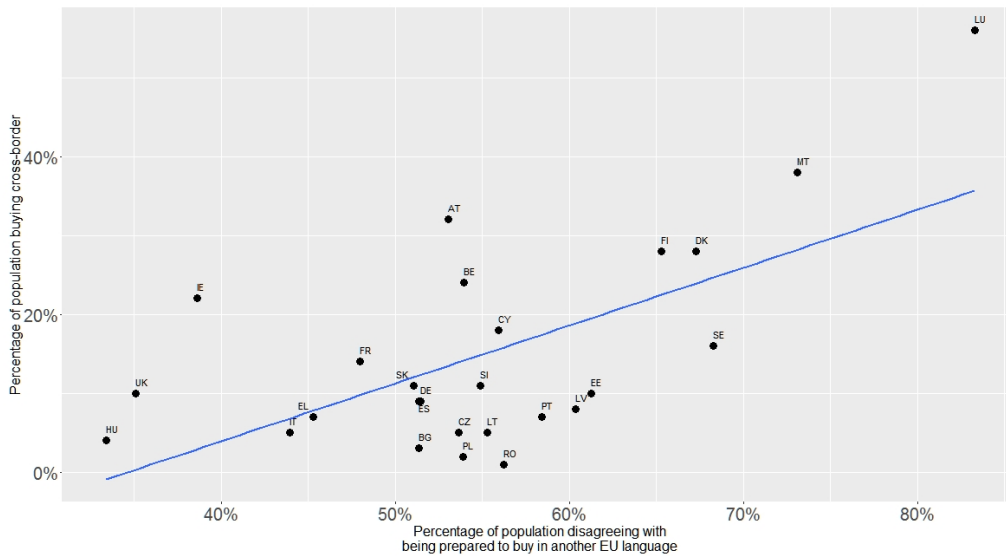
Figure 24: Cross-border language barriers for consumers by country

Source: Compiled by the authors based on (European Commission, 2012d)

²⁵ Companies with 10 or more employees, excluding the financial sector, are considered

Although the figure is worrisome by itself, the language barrier is likely to be bigger than perceived by consumers. Our thesis is that the language barrier is, in fact, a hidden barrier for many on-line consumers. In analysing for each country the relationship between the percentage of population buying on-line and the percentage of population who consider that they are not prepared to buy in another EU language, we have found that there is a positive correlation between the percentage of population buying on-line to other countries and the perception of the language barrier as shown in Figure 25. It is quite surprising that in those countries whose population tend to buy in other countries, people are more likely to think that they are not prepared to buy in a foreign language.

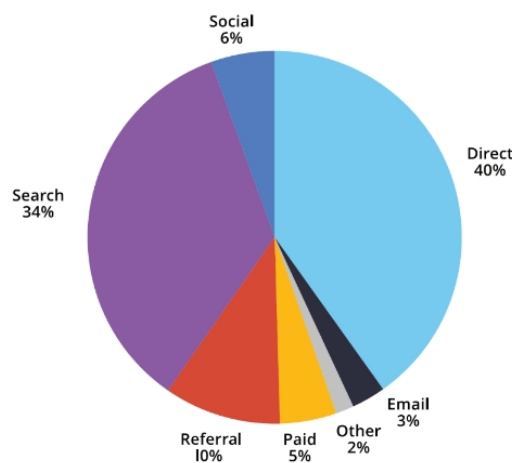
Figure 25: Relationship between language barrier and percentage of buyers



Source: Compiled by the authors based on (European Commission, 2012c, 2014a)

People are not aware of what they do not know. The three main sources of e-commerce traffic, namely direct traffic (40 %), organic traffic from search engines (34 %), and referral traffic (10 %), account for 84 % of e-commerce traffic as seen in Figure 26, and this traffic mainly comes from pages in the language of the user. Users seldom arrive to e-commerce pages in languages different from their own and therefore remain unaware of the problem.

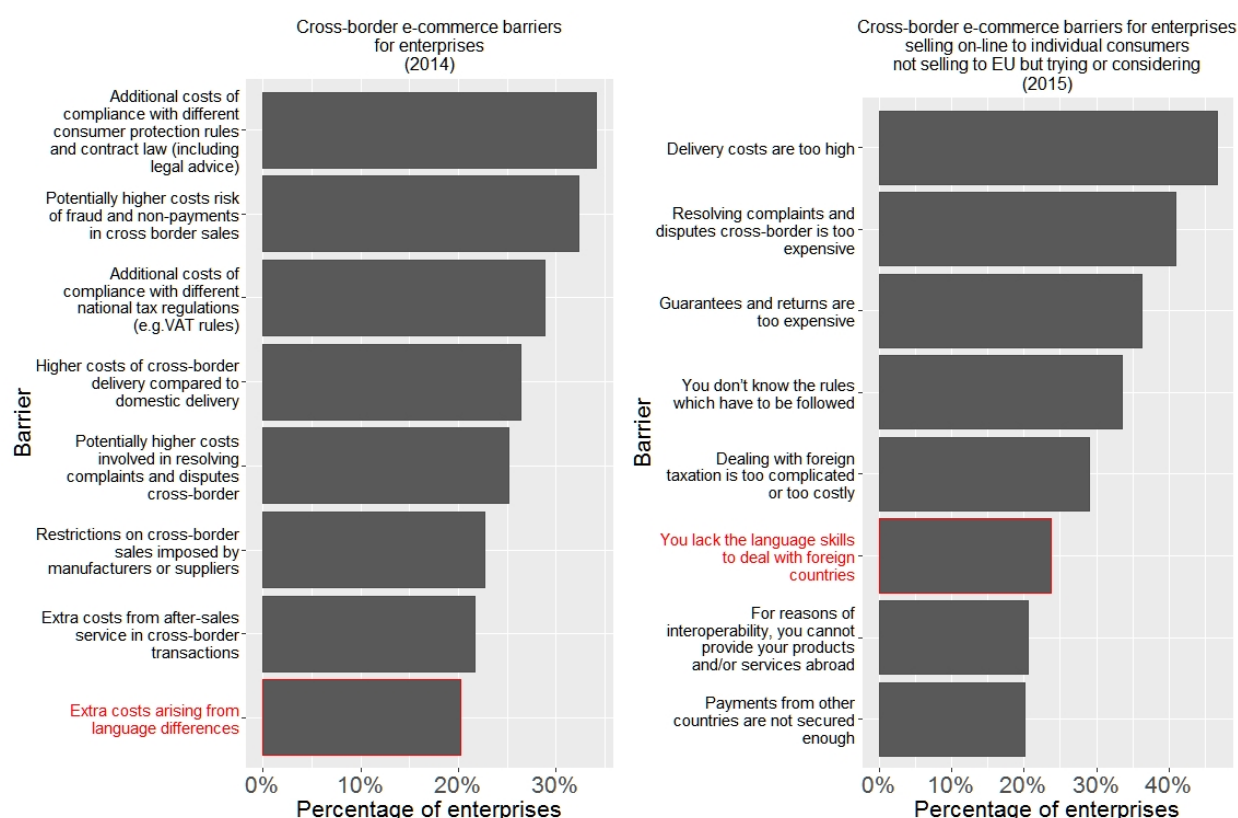
Figure 26: E-commerce traffic broken down by source



Source: Figure of Yotpo's global user database of over 120.000 online businesses (YOTPO, 2015)

In a recent study commissioned by Ecommerce Europe, the main European association of online vendors, the language differences across the European Union are also seen as a difficult barrier to overcome for 29 % of companies that sell cross-borders (Ecommerce Foundation, 2016). Analysing Eurostat data, we reach a similar conclusion. Close to 30 % of the companies who sell on-line to individual consumers consider that language skills are an important barrier as can be seen in Figure 27. Still, companies consider that there are other important barriers, such as a defragmented taxation and regulatory environment and a still inadequate pan-European distribution service.

Figure 27: Cross-border e-commerce barriers for enterprises selling on-line to individual consumers not selling to EU but trying or considering (percentage of companies)

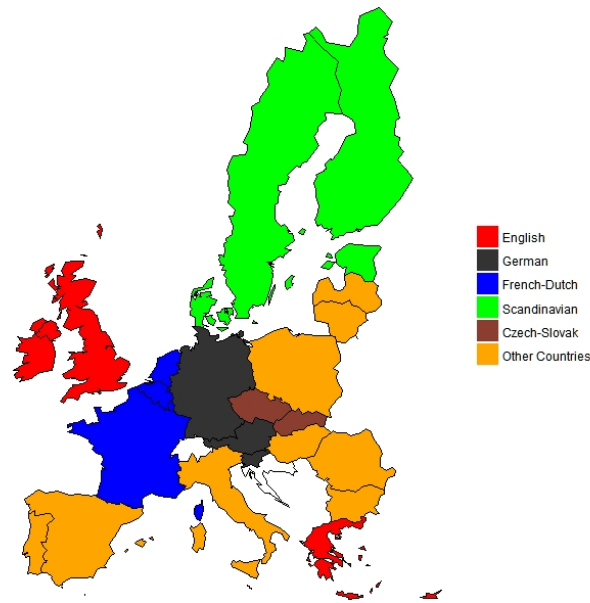


Source: Compiled by the authors based on (European Commission, 2015b, 2015c)

3.2.5.3 A fragmented DSM

Under these circumstances it is unlikely that the DSM works as an integrated market. Pavel (2010) suggests that the fragmentation of e-commerce in Europe is likely to depend on language barriers, differences in regulation and on specific consumer preferences. By performing a cluster analysis of the percentage of e-commerce users buying from other countries we have found a strong pattern of six clusters of countries mainly shaped by culture and language similarities²⁶. Most of the countries within the clusters share the same or closely related languages as shown in Figure 28.

²⁶ The analysis is detailed in Annex 7.4.1.

Figure 28: Clusters of cross-border e-commerce (consumers' side)

Source: Compiled by the authors based on (Civic Consulting, 2011)

The cross-border e-commerce between countries of the different clusters is very scarce, except when the destination country is one of the biggest economies. Additionally, the group “Other Countries” that includes most of the countries with smaller languages in Europe, show very low cross-border e-commerce, both to countries within the same group and to countries in other clusters, except to the big economies. These countries remain isolated and clearly disadvantaged because no European on-line buyers from other countries tend to buy in these countries, while e-shoppers in these countries face difficulties to buy in other EU countries. In considering the supply side (web merchants selling to other EU-countries) we perform a similar analysis as described in Annex 7.4.2 to find a similar pattern of clusters. The analysis suggests that language and cultural barriers are hindering a truly integrated DSM both for suppliers and consumers. However, there are other commercial reasons that hamper the development of the DSM, as the geo-blocking that online traders use to segment the market based on customers’ residence. The European Commission has issued a proposal to prohibit this discrimination in order to boost the DSM (European Commission, 2016h).

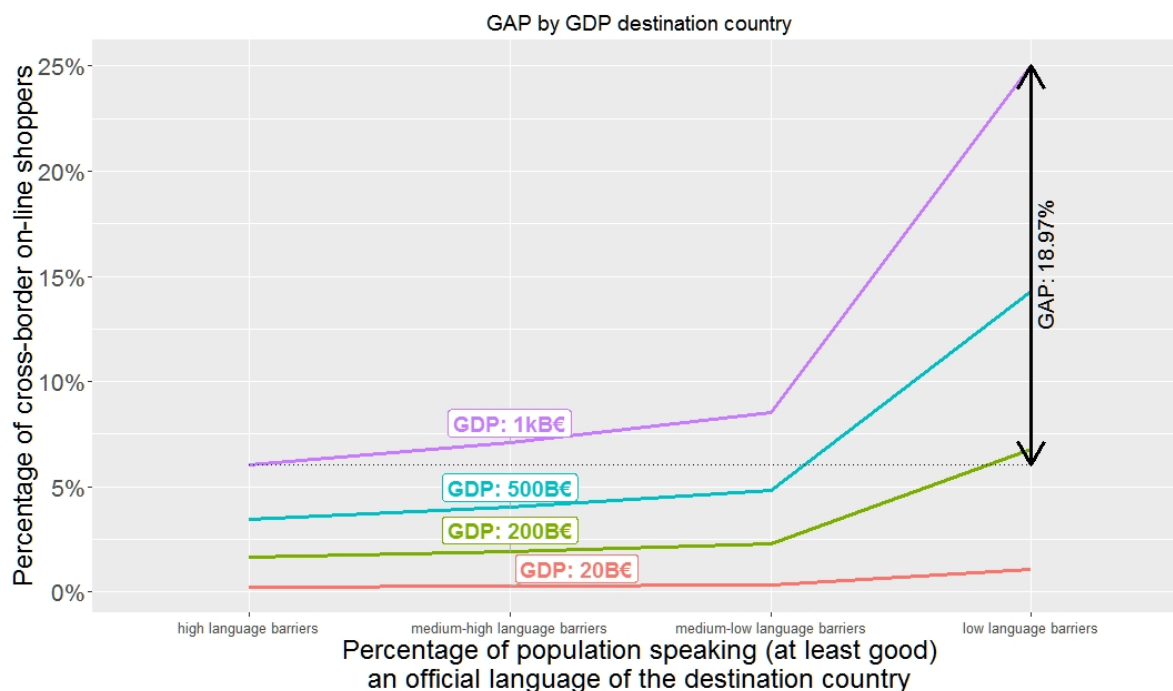
3.2.5.4 Assessing the effect of language barriers on cross-border e-commerce

To quantify the language barriers, we have made several regression analyses²⁷ to find that having low language barriers increases 142 % the number of on-line shoppers buying cross-border compared to having high language barriers. It means that, on average, the number of cross-border shoppers is about four-fold higher between countries with low language barriers compared to high language barriers.

²⁷ Annexes 7.5.2 (demand side) and 7.5.3 (supply side).

The results can be shown in a simple graphic presentation in Figure 29²⁸. Take a destination country with GDP of 1 000 billion euro. The language gap is close to 19 %, increasing, on average, the percentage of on-line shoppers buying from a foreign country from 6 % to 25 %. The results suggest that language is an important and relevant barrier for consumers to build a truly integrated DSM.

Figure 29: Descriptive regression results for cross-border on-line shoppers



Source: Compiled by the authors based on (Civic Consulting, 2011; European Commission, 2014a; EUROSTAT, 2016c)

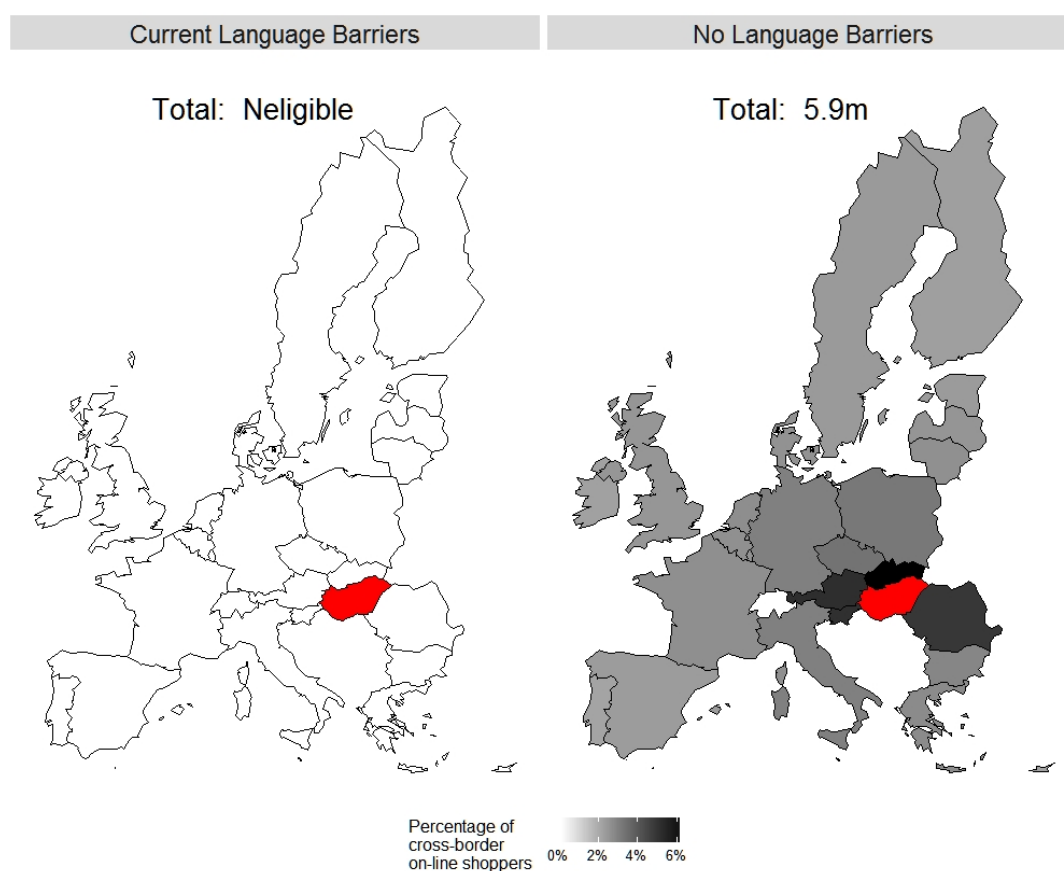
For retailers we find that, although the language barrier still plays a role, logistics seems to have a more important effect (both selling to neighbouring countries and to closer countries). This is in line with the barriers described in Figure 27.

Using these results we have simulated what would happen in a theoretical scenario if high quality HLT could allow a simple and efficient mean to automatic translation of the e-commerce websites²⁹. Take Hungary, a country with a linguistically isolated language and a high percentage of population not speaking any foreign language. In the current situation, the EU population shopping on-line from other countries to Hungary is negligible, while we estimate that overcoming language barriers will increase the population from other countries buying on line from Hungary up to 5.9 million people as shown in Figure 30. That is more than twice the current population buying on-line in Hungary (2.2 million).

²⁸ For neighbouring countries with the average values of distance, the average value of regulatory quality of the destination country, and four different economic sizes (GDP ranging from 20 billion euro to 1 000 billion euro). Destination country different than Germany, UK and France.

²⁹ The detailed results can be found in Annexes 7.5.2.6 (demand side) and 7.5.3.6 (supply side).

Figure 30: Population shopping on-line in Hungary from other countries depending on language barriers



Source: Compiled by the authors based on (Civic Consulting, 2011; European Commission, 2014a; EUROSTAT, 2016c)

3.2.5.5 Effects on the European economy of an integrated DSM

The key question becomes what would be the overall effect of a truly integrated DSM on the welfare of the European society and its economy. It is expected that the increased on-line offering will foster a more competitive market that will promote price competition. On-line prices will lower, yielding an increased consumer welfare. The European Consumer Centres Network has estimated that European consumers could save 11.7 billion euro annually when shopping online if they could choose from the whole range of goods and services within the EU28 (European Commission, 2015f). Moreover, in a hypothetical scenario of a 15 % share of Internet retailing in total retail – currently internet shopping accounts for 6.4 % of all retail purchases in Europe (Hunter & Wilson, 2015) – and a true single EU consumer market, it was estimated that total welfare gains for EU consumers due to the increased competition could reach 204.5 billion euro annually (1.7 % of EU GDP), based on 2011 data. That is four times higher compared to the current fragmented market (Civic Consulting, 2011).

More interesting is to assess the effect on the whole economy. The macro-economic effects of reducing the barriers to e-commerce has been estimated by Cardona, Duch-Brown, Francois, Martens, & Yang (2015) by using a macro-economic general equilibrium model. Their conclusions are that a higher level of transactions between countries increase price competition and therefore has a negative effect on the domestic retail markets (-2.6 %). However, this increased competition has a positive effect in other sectors (up to 2.6 %) and on household consumption (1.7 %). They suggest that increasing cross-border e-commerce brings gains in efficiency and welfare distribution from retailers and households to other sectors while having a moderate effect on EU GDP (0.14 % increase).

In the current absence of technologies that would allow overcoming language barriers in a seamless, efficient and speedy way we see that the situation is not improving and the percentage of consumers and retailers buying and selling on-line show symptoms of stagnation. Our results suggest that having effective HLT could improve the situation of cross-border e-commerce by effectively tackling the problems associated with language barriers. However, it is very important to consider not only large languages, but also European smaller languages, in order not to leave behind citizens who are using those languages. On the other hand, effective HLT for all languages – large and small – will make households and businesses in these countries particularly better-off, thus fostering a truly integrated and fair European DSM.

3.2.6 Impact on SMEs

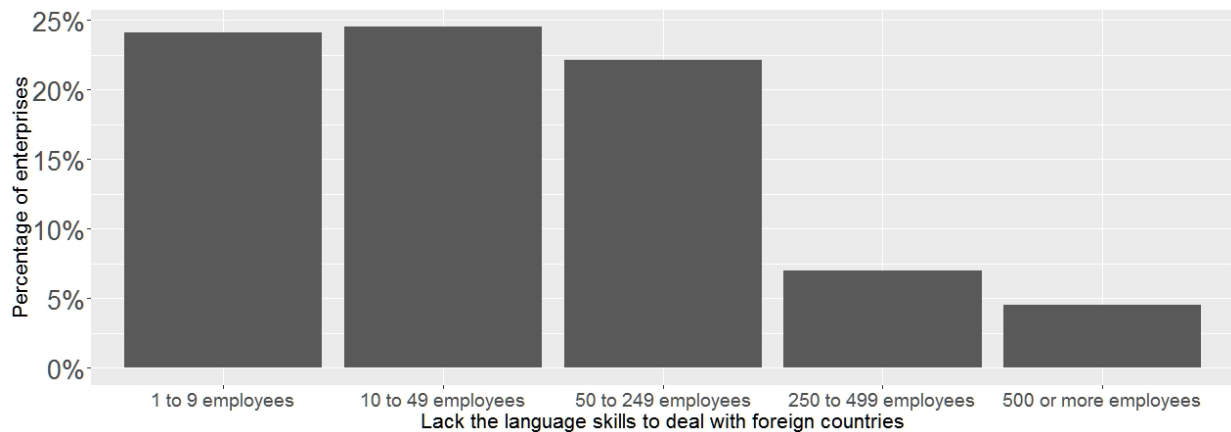
SMEs are a crucial pillar of the European economy³⁰ and the effect of not having adequate language skills within the SMEs can be particularly troublesome for Europe. Different factors, such as increased exposure to technical know-how, higher market-awareness and improved efficiency, are considered key in order to increase the productivity of exporting SMEs by as much as 3.7 % compared to the industry mean (Hagen, Foreman-Peck, Davila-Philippon, Nordgren, & Hagen, 2006). However, the National Centre for Languages (Hagen et al., 2006) estimated³¹ that at least 945 000 European SMEs were losing business due to lack of language skills, with an average loss per business over a three year period of 325 000 euro. Longer-term business partnerships depend upon relationship-building and relationship-management, making cultural and linguistic knowledge of the target country an essential ingredient. For SMEs involved in cross-border trade, the level of preparedness in the use of languages for publishing content is extremely important, as the number of languages must increase. Although currently 60 % of SMEs are capable of publishing in more than one language – except for the UK, where 60 % of SMEs are monolingual –, language barriers are more significant for SMEs compared to larger companies as shown in Figure 31. Therefore, the cost of translation becomes the main barrier to implementing an online multilingual content strategy, as publishing costs rise by one third.

Retail, media and manufacturing are the leading markets today driving the adoption of LT and multilingual content publishing when going cross-border in Europe. In particular, wholesale and retail trade is the most important sector for micro and small SMEs in terms of employment, value added and number of enterprises (Muller et al., 2014), and therefore it is one of the sectors that could benefit more from cross-border transactions. However, for SMEs, selling on-line abroad is more challenging than for larger companies when considering language barriers, as can be seen in Figure 32. In 2015, 23 % of larger companies sold through electronic channels to other EU countries compared to only 12 % and 7 % of medium and small companies respectively.

³⁰ SMEs represent around 99.8 % of all enterprises and account for around two-thirds of total employment in 2014 (71.4 % of the increase in employment in that year) in the non-financial business sector (Muller et al., 2014).

³¹ Based on a survey to 20.000 SMEs promoted by the Center for Information on Language Teaching (CILT).

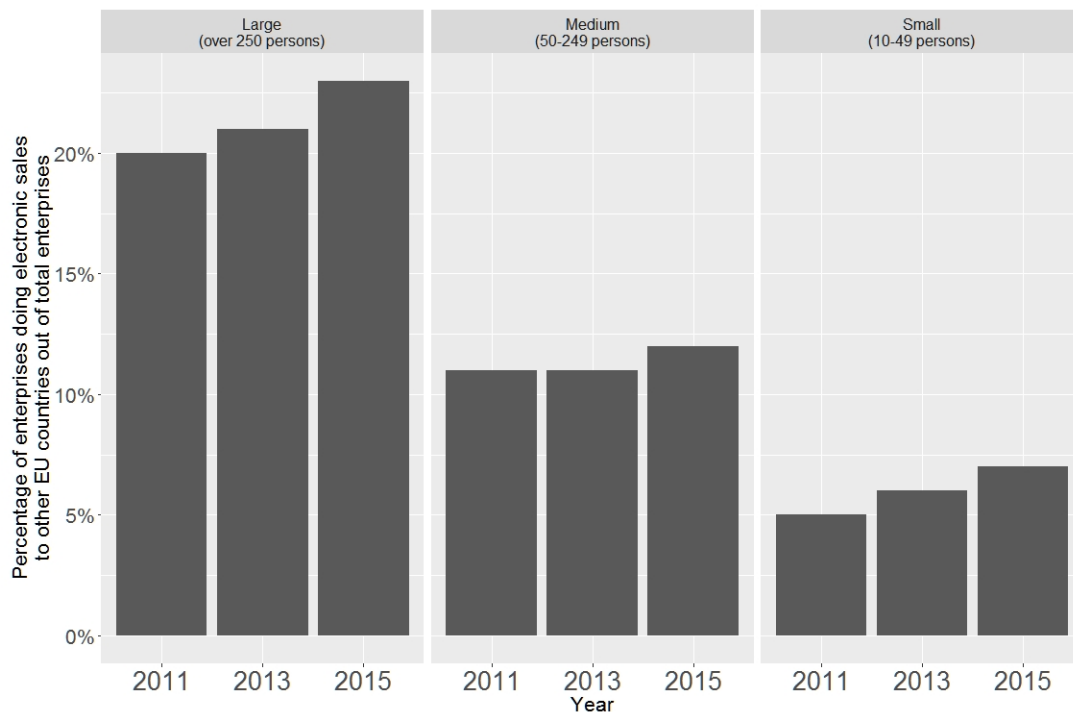
Figure 31: Cross-border e-commerce language barrier for enterprises selling on-line to individual consumers that are trying or considering to sell abroad by size



Source: Compiled by the authors based on (European Commission, 2015c)

One of the reasons of this gap is the language barrier. Using the Community Survey on ICT Usage and e-Commerce of 2009, Pavel (2010) found that the only highly significant barriers to electronic sales between large and small and medium enterprises were language problems and technical issues. Regarding the language barriers, it is expected that efficient HLT solutions will help to foster an increasingly level playing field where small European retailers could compete more easily with big web-stores, many of them in the hands of non-European companies.

Figure 32: Companies doing electronic sales to other EU countries by size



Source: Compiled by the authors based on (EUROSTAT, 2016b)

Regarding the technical issues, one of the last ICT innovations that can help SMEs to increase their productivity and competitiveness is the cloud computing paradigm. However, the lack of cloud solutions in local languages might hinder the adoption of these technologies by SMEs in the short-term (Bradshaw, D., Folco, G., Cattaneo, G., & Kolding, M., 2012).

Current language strategies of SMEs are based on traditional approaches such as recruiting native speakers (22 % of firms), website translations (50 %), using translators, and hiring local agents. Investing in those elements can increase the export sales proportion up to 44.5 %. However, not all SMEs are ready to invest on those capabilities. High quality HLT service adapted to the specificities of the SMEs could be the answer by providing feasible, quick and efficient solutions to these companies.

3.2.7 Other effects of multilingualism on cross-border trade and businesses in Europe

In the business sphere, not only the e-commerce services are facing the challenges derived from a multilingual economy. Multilingualism, one of the cultural cornerstones of Europe, is also one of the main obstacles of a truly connected, language-border crossing Single Market. Linguistic interoperability is vital for European economic growth and competitiveness. Significant amounts of business are being lost by European enterprises due to lack of language skills and the capacity to contextualise information. Even the cradle of the main language for business, the United Kingdom, has recognized the enormous cost of language barriers in cross-border trading. According to a study commissioned by the United Kingdom Trade & Investment Department (replaced by the Department for International Trade in July 2016), the cost of lacking language skills in international trading in the UK could be estimated as 3.5 % of its national income in 2006, or £48 billion. Moreover, these costs seems to be increasing, probably due to the emergence of trades with developing countries like China, where English skills are not enough (Foreman-Peck, J. & Wang, Y., 2013). An English-only strategy may fall short. For instance, in eastern European countries, such as Lithuania, businesses are likely to demand Russian skills (British Council, 2011).

Although the impact of not having language skills on international trade is likely to be very large, the perception of language barriers for cross-border trade follows a pattern similar to the perception of barriers for consumers in cross-border e-commerce. Again, these barriers are likely to be hidden from businesses. For instance, the businesses that are aware of cultural difficulties in the UK are those with higher export intensity, while companies claiming that they do not experience cultural barriers tend to show lower levels of exports and lower language skills (Foreman-Peck, J. & Wang, Y., 2013). In other countries, such as Romania, there is also a worrying lack of awareness of the importance of speaking other partners' languages (British Council, 2011). This makes it particularly challenging to effectively develop traditional language policies.

Investing in language skills is therefore expected to yield a high return although it is not always feasible for companies. HLT could provide cost-effective solutions to compensate the deficit of language skills in making businesses, thereby opening new opportunities for international trade for European companies.

3.3 Human language technologies and public policies

This chapter analyses how the EU is facing multilingualism and HLT in current ICT policies and in the digital agenda. We also assess the experience in other multilingual societies such as South Africa or India. Eventually, we analyse the policy recommendations of the industry and research groups in the EU.

When considering multilingualism policies, there is always a trade-off between effectiveness and fairness, between utopia and reality, and between the preservation of cultural heritage and diversity

and fostering an effective integrated global market. Multilingual policies then become an uncomfortable topic. On the one hand, preserving multilingualism is strongly rooted to the essence of European values of cultural diversity and it is likely to be the only way to create a fair and truly integrated European Union while preserving our culture. On the other hand, creating a truly integrated Europe will only be possible if current language barriers are overcome, but the cost seems to be unfeasible. Fostering policy measures aimed at spreading proficiency in a lingua franca such as English could be efficient but will yield negative cultural and political effects, such as creating a larger language gap, substantially reshaping our thinking and behaviour and reducing the willingness of EU citizens to maintain their languages (Gazzola, M., 2014).

There are currently 24 EU official languages. EU citizens have the right to use any of these languages in contacting with the EU institutions. EU regulations and other legislative texts are published in all official languages and the EU provides general information about its policies in all its official languages. However, more specialised content is only provided in the most widely spoken EU languages. Regarding other regional and smaller languages, the EC encourages linguistic diversity to the highest extent possible, although the legal status and support for these languages is determined by national governments (European Commission, 2016c). Due to the different needs of national and regional languages, two different organisations cater for those needs in Europe: EFNIL for the national languages and NPLD for the regional ones.

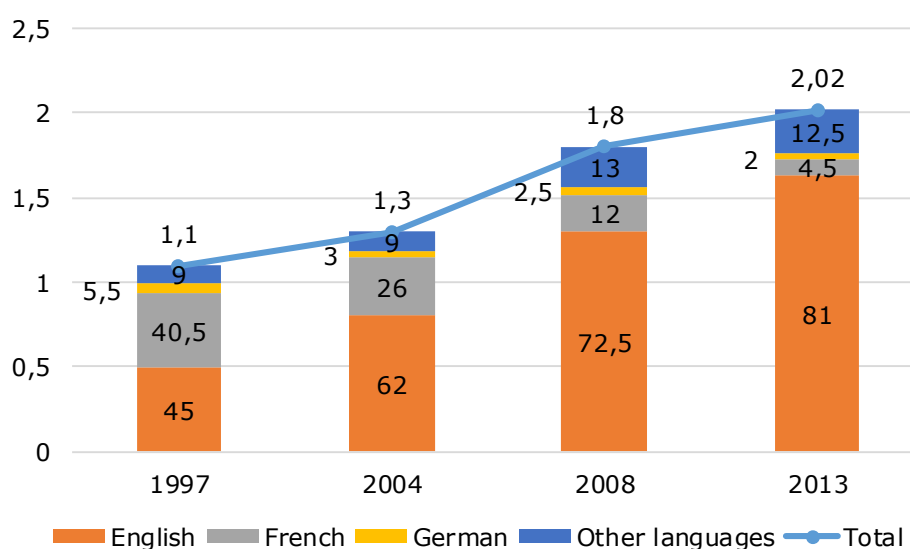
Formed in 2003, the European Federation of National Institutions for Language (EFNIL, 2016) has institutional members from 30 countries. Its role includes monitoring the official language(s) of their country, advising on language use or developing language policy. It provides a forum for these institutions to exchange information about their work and to gather and publish information about language use and policy within the EU. EFNIL encourages the study of the official EU languages and a coordinated approach towards mother-tongue and foreign language learning as a means of promoting linguistic and cultural diversity within the European Union. There is an increased awareness among EFNIL members of the relevance and importance of HLT on several counts. First, as a vital component and a requirement for the sustainability of their respective national languages in the digital age. Second, as a research and productivity tool that has an increasing impact on their daily work. Third, EFNIL members, many representing the central academic institutions for their language, can contribute to the technology support for their language through the invaluable language resources they have developed. As an example, EFNIL is running a pilot project (EFNILEX) aimed at developing LT support for the production of bilingual dictionaries between language pairs which are considered by mainstream publishing houses as commercially unviable.

On their side, the Network to Promote Linguistic Diversity (NPLD, 2016) is a pan-European network which works with constitutional, regional and smaller state languages. NPLD was established in 2007 and has already asserted itself as the main voice of those linguistic communities that are not the official languages of the EU. NPLD has two main goals: the first is to take advantage of the growth in knowledge and expertise which is now available in the area of language reinforcement by ensuring that it is shared. This is done mainly through meetings and seminars, and is in the process of being further developed through the expansion of a digital library on language planning for its members. The second goal concerns the issue of policy development at a European level. Although much is said by the European institutions about the importance of linguistic diversity, very few policy initiatives are undertaken and even less funding is provided to support European linguistic diversity. They aim to highlight this deficiency and to promote the need for more support for all indigenous languages of Europe to ensure that our rich landscape of languages, many of them highly endangered, survive into the future. ICT and social media will play a vital role in the future survival of most, if not all of the languages of Europe (NPLD & EFNIL, 2016).

However, the EC itself has moved towards a monolingual regime within their institutions and, in practice, English has become the only working language in the EC. Taking the evolution of the number

of pages translated by source language from 1997 to 2013, we observe that, while in 1997 English, French and to some extent German were the usual working languages, in 2013, English was by far the most used language as seen in Figure 33. Regarding the target language of pages translated, English was again, by far, the most relevant language, with around 250,000 pages, followed by French and German (between 100 000 and 150 000 pages translated) and the rest of languages (less than 100 000 pages) (European Commission, 2014c). Both indicators (source language and target language of pages translated by DG Translation) places English as the dominant language in the European institutions. This dominance cannot be justified in terms of fairness, although it is likely to be more realistic, efficient and feasible.

Figure 33: Evolution of number of pages translated by source language (percentage of pages; total in millions)



Source: (European Commission, 2014c)

3.3.1 Trends in multilingualism and HLT in the EU bodies

In analysing current standing of European bodies regarding multilingualism we focus on the European Parliament and the DG connect, and particularly on the Digital Single Market, because the main goal of this report is to understand the role of HLT in the digital age. To do so we used 3 different sources:

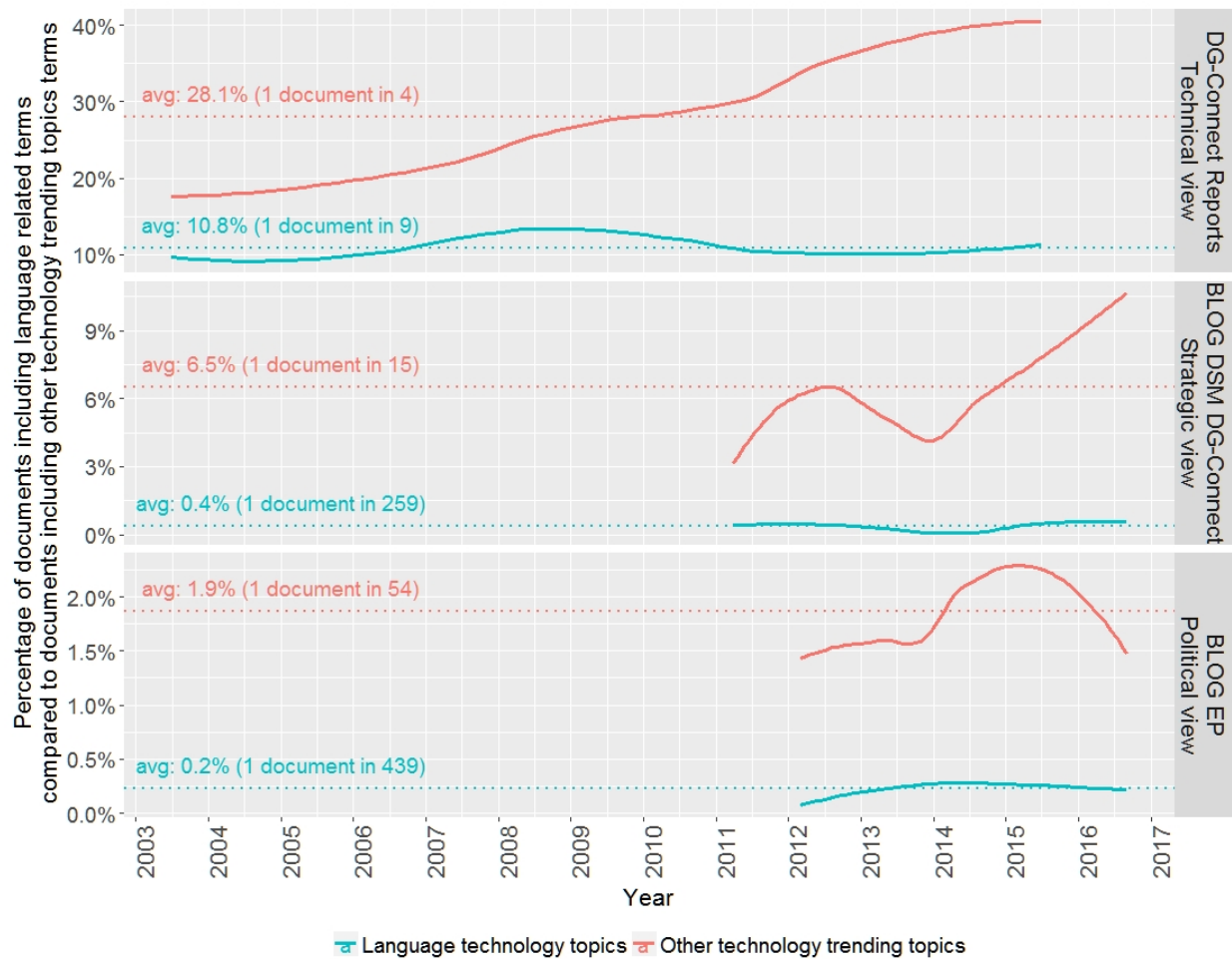
- Technical documents of the DG Connect from the inventory of the reports on the studies completed by the European Commission Directorate General for Communications, Networks, Content and Technology (478 documents from 2003 to 2015) (DG CONNECT, 2014)
- Blog of the Digital Single Market of the DG Connect (440 posts from 06/04/2011 to 21/09/2016) (European Commission, 2015e)
- Blogs of the European Parliament (2.006 posts from 02/03/2012 to 28/09/2016) (European Parliament, 2015)

These sources provide a good overview of the political, strategic, and technical point of view of the European institutions. The analysis is made by using text-mining techniques that allow looking for specific terms and how these terms relate to each other. We have selected different terms related to human language technologies and we have made a benchmark comparing those terms to other terms

related to trending topics in the digital ecosystem. The terms are listed in Annex 7.6. We have assessed whether language technologies are considered a relevant topic and how the perception is evolving by analysing the percentage of documents that include terms related to a language topic compared to the percentage of documents that include terms related to other trending technology topics out of the total number of documents. We have further assessed how deeply each of the concepts is covered by analysing the average density of sentences per document for each set of the terms. Acknowledging the limitations of the analysis, we still think that they give an overall good idea about the differences between multilingualism and HLT topics compared to other trending technology topics in this relevant source of information.

The reports of the DG Connect provide the technical vision, the Digital Single Market Blog of the DG Connect is related to the strategic opinion about the DSM and its most relevant issues, while the blog of the EP represent the political view. The results of the analysis are shown in Figure 34 and Figure 35. Regarding the technical view, on average 28 % of the documents include references to trending technology topics while only 11 % of the documents include references to language technology topics and the gap has been widening since 2003. This suggests that language technologies are increasingly becoming a less important technology topic, although it was well covered compared to other technology topics 10 years ago. In analysing the strategic standing of the DSM, we have found that language technology is an irrelevant topic (only 1 post in 242 are related to the topic, compared to 1 post in 15 of other technology topics) and the gap with other trending technology topics is substantially increasing. This is quite surprising considering that language barriers, which could be overcome by using these technologies, are one of the main challenges to creating a truly integrated and fair DSM as we have seen in the impact analysis. Regarding the political point of view reflected in the blog of the EP, we also see that language technology topics have very low relevance. At least the gap with other trending topics is not widening.

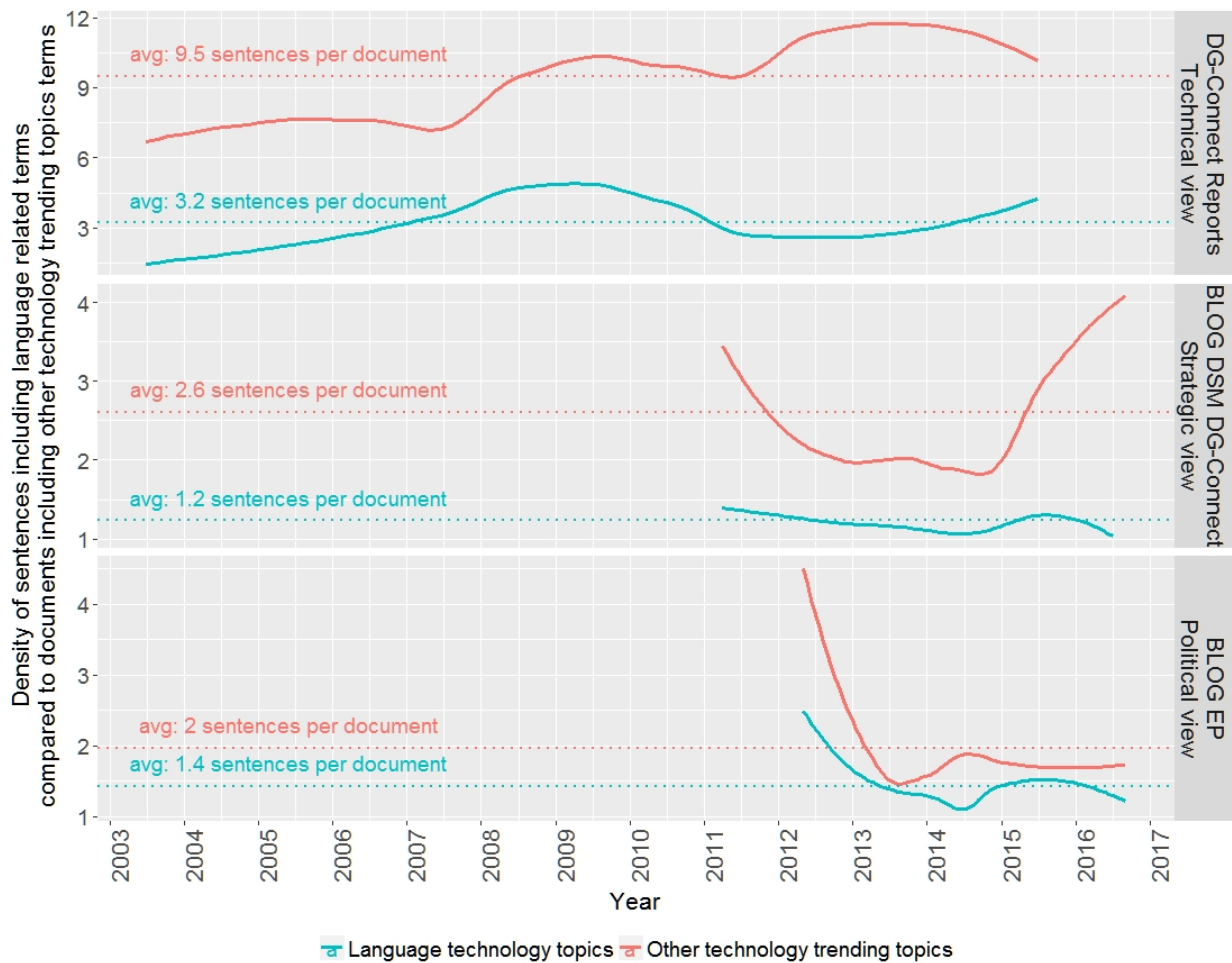
Figure 34: Average percentage of documents including the topics (language technology topics compared to other trending technology topics)



Source: Compiled by the authors based on (DG CONNECT, 2014; European Commission, 2015e; European Parliament, 2015)

In analysing how deeply the different topics are covered in the documents we find a similar trend. In the reports of the DG Connect, the gap between language technology topics and other technology topics is quite big (3.2 sentences per document on average compared to 9.5, respectively), and the gap is growing slightly. This suggests not only that the topic is being included in fewer documents but also that language technology topics are being examined in less detail. Most worrying is the trend for the DSM strategy. Although the gap is smaller, it has substantially increased in the last months. This confirms that language technologies have little relevance for the DSM. Regarding the political vision, the gap is small and has not changed in the last few years. Although there is a much lower number of EP blog posts including these topics, at least when they are included, it seems that they are analysed with a good level of detail.

Figure 35: Average number of sentences per document including the topic (language technology topics compared to other trending technology topics)



Source: Compiled by the authors based on (DG CONNECT, 2014; European Commission, 2015e; European Parliament, 2015)

3.3.2 EU multilingualism and HLT policies

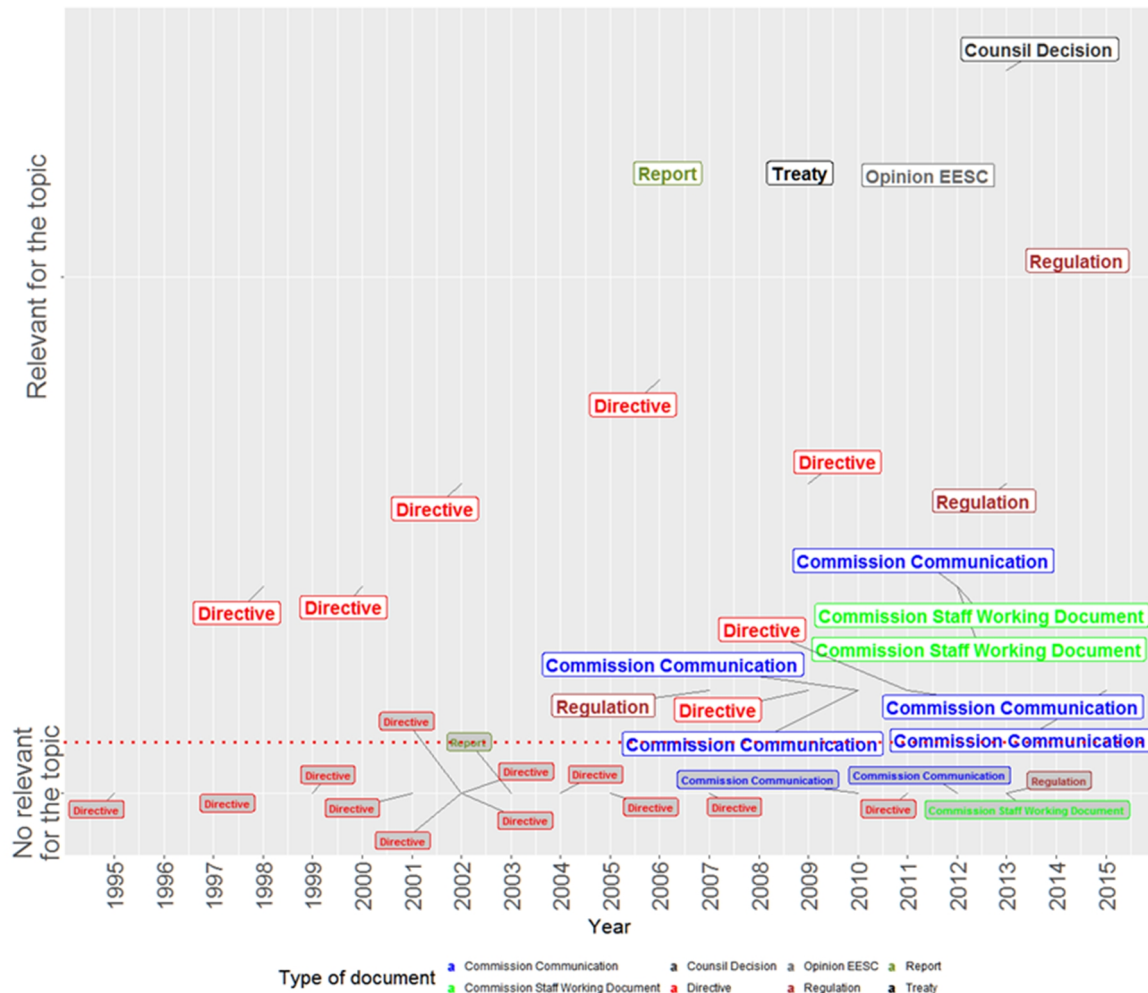
To assess the standing of the EU in relation to HLT and multilingualism we have analysed 38 relevant policy and regulatory documents³² related to public services, the digital economy and the EU single market since 1995. The list of documents analysed is shown in Figure 58 of Annex 7.6.2. These documents include EU directives (19), Commission communications (7), one Council decision, regulations (4), one opinion of the European Economic and Social Committee, reports about the directives (2) and Commission staff working documents (3). Although we acknowledge that the results may not give a totally objective and comprehensive overview, we still think that they provide a good indicator of the role of multilingualism and HLT on main European ICT policies.

Before 2006 we only found references to terms related to “linguistic diversity” in several directives as can be seen in Figure 36. For instance, the Directive 2000/31/EC on electronic commerce states that “[...]”

³² These documents have been selected based on the desk research made in the project.

it should not hinder measures which Member States might adopt in conformity with Community law to achieve social, cultural and democratic goals taking into account their linguistic diversity’.

Figure 36: Official EU documents and their relevance to the topic based on number of sentences by type of document and year



Source: Compiled by the authors

In white background those documents including terms related to Multilingualism

In the Directive 2006/123/EC, services in the internal market (European Commission, 2006) appear to need to promote national languages: ‘[one of the overriding reasons relating to the public interest] is the promotion of the national language’. There is another brief reference to language barriers in the Directive 2011/24/EU on the Application of patient’s rights in cross-border healthcare: ‘It should be noted that the impact on national health systems caused by patient mobility might vary between Member States or between regions within a Member State, depending on factors such as geographical location, language barriers, location of hospitals in border regions or the size of the population and healthcare budget’.

There are some interesting references to the effect of language barriers in cross-border e-commerce (based on different surveys) in the report “Study on the economic impact of the Electronic Commerce Directive” (Kastberg et al., 2007) and in the Commission staff working document “Bringing e-commerce benefits to consumers” (European Commission, 2012a).

In the Commission communication document “A Digital Agenda for Europe” (European Commission, 2010) there is a brief mention to multilingual contents and services: ‘Work with stakeholders to develop a new generation of web-based applications and services, including for multilingual content and services, by supporting standards and open platforms through EU-funded programmes’.

However it is the opinion of the European Economic and Social Committee on A Digital Agenda for Europe (European Economic and Social Committee, 2011) that goes to the point by making some quite interesting comments: (1) ‘Investment should be targeted at finding innovative solutions to the challenges caused by language diversity in the EU’; (2) ‘Language diversity is a special challenge for the 2020 vision’; and (3) ‘Language diversity is a special challenge for Europe when trying to create a vibrant single market for online goods and services’.

In the Commission communication, “Single Market Act, twelve levers to boost growth and strengthen confidence, Working together to create new growth” (European Commission, 2011), the only mention to HLT regards to automatic translation technologies that ‘should be developed, in order to facilitate exchanges between authorities’. It is neither the citizens nor the markets, but the possibility of national authorities and the Commission working more closely is what matters.

In the Commission communication “A coherent framework for building trust in the Digital Single Market for e-commerce and online services” (European Commission, 2012b) there is a reference to multilingualism and price comparison sites. Although it is a brief mention it points out the relevance of language barriers for one of the sources of e-commerce traffic: ‘[price comparison] sites could usefully offer citizens a wider choice through multilingual, cross-border information, thereby stimulating domestic competition and helping to build the Single Market’.

In the Commission communication “A Digital Single Market Strategy for Europe” (European Commission, 2015a) there is a strong support for multilingualism in the DSM although there are no further mentions to HLT in the document: ‘The Commission aims to support an inclusive Digital Single Market in which citizens and businesses have the necessary skills and can benefit from interlinked and multi-lingual e-services, from e-government, e-justice, e-health, e-energy or e-transport’. In fact, the document raises serious concerns about legal, administrative, copyright, and logistic barriers for the DSM but does not make any mention to language barriers that are likely to be one of the most challenging topics to the effective construction of the DSM.

The actions suggested for language technologies within the Digital Single Market strategy show a higher interest on HLT as a relevant tool for Europe to face the challenges of multilingualism and are focused on two levels (European Commission, 2016e): on the one hand research and innovation and on the other creating a complete infrastructure for language resources and tools. Regarding research actions, the “Council Decision of 3 December 2013 establishing the specific programme implementing Horizon 2020” (European Commission, 2013a) mentioned several times language, big data, and machine learning technologies: (1) ‘The issues of multilingualism, translation and circulation of ideas across Europe and from and to Europe and how they form part of a common European intellectual heritage will be explored’; (2) ‘This includes new technologies for arts, language, learning, interaction, digital preservation, web design, content access, analytics and media; and intelligent and adaptive information management systems based on advanced data mining, machine learning, statistical analysis and visual computing technologies’; (3) ‘It will do so by providing professionals and citizens with new tools to create, access, exploit, preserve and reuse all forms of digital content in any language and to model, analyse and visualise vast amounts of data (big data), including linked data’. Within H2020, research and innovation in machine translation was supported, including HLT topic ICT 17 – 2014: Cracking the language barrier (European Commission, 2013d) and also in the context of big data – e.g., enabling big data applications in multiple calls. Regarding the infrastructure, the “Guidelines for trans-European networks in the area of telecommunications infrastructure” (European Commission, 2014b) included the main topics for common projects of interest within the Connecting Europe Facility (CEF) and it is interesting that there are a couple of topics related to HLT: (1) ‘Automated translation: this refers to

machine-translation engines and specialised language resources including the necessary tools and programming interfaces needed to operate pan-European digital services in a multilingual environment’; (2) ‘This refers to a platform for the single access point to multilingual (official languages of the institutions of the Union) datasets held by public bodies in the Union at European, national, regional and local levels; query and visualisation tools of the data sets; assurance that the available datasets are properly anonymised, licensed and where applicable priced to be published, redistributed and reused, including a data provenance audit trail’.

The main conclusion of the analysis is that language technologies have not been considered one of the most relevant issues for the EU, although there are 24 official languages and more than 60 regional and smaller languages. In fact, the Policy Recommendations for the Promotion of Multilingualism in the European Union of the Civil Society Platform on Multilingualism (2011)³³ only included four groups of policies, namely planning policies, language diversity and social inclusion policies, education policies, and translation policies without making any mention to the role of HLT to effectively tackle linguistic barriers in Europe. Language technologies were mainly considered as a tool to improve learning and teaching of foreign languages or to translate live performances in the arts and cultures sector.

The situation shifted in 2011, when the European Economic and Social Committee raised concerns about the challenges of European language diversity for the 2020 vision. However, the Digital Single Market Strategy of 2015 only makes a brief reference to multilingual services. There is never any mention about the role of HLT in providing these services, nor mention of the role of multilingualism as one of the most important barriers for the EU DSM. The good news is that within the DSM strategy there are two actions that are oriented to promote the research on HLT and on providing these technologies as a service. The bad news is that the levels of investments on HLT seem to be much lower than one would expect in a technology that is likely to be crucial to foster a vibrant, fair and integrated EU. In fact, after analysing the investments in HLT in the Horizon 2020 program we found the following figures (Rossi, 2016):

- ICT-17 (15 million euro): Cracking the language barrier.
- Other projects in the big data topic (around 6 million euro).
- CEF-AT in 2014 WP (4 million euro)
- CEF-AT in 2015 WP (8 million euro)
- CEF-AT in 2016 WP (9.5 million euro)

On top of that, the ICT Policy Support Programme (part of the Competitiveness and Innovation Framework Programme) has invested 29.6 million euro. Regarding previous investments, the estimated total EU contribution to HLT related projects in the 7th Framework Program³⁴ from 2008 to 2014 was about 160.9 million euro³⁵, making a yearly average of 26.8 million euro. Detailed information regarding the projects is included in Annex 7.7.

These amounts seem too small to have a substantial impact as suggested by some of the experts that have participated in this study. We can compare those figures to the estimated 500 million euro that Google spent in 2014 to acquire DeepMind technologies, a London-based artificial intelligence firm (The

³³ The ECSPM was launched by the EU as the main tool in the context of multilingualism to improve the policy process (M. Prys Jones, 2013) and re-launched in 2012. Currently is an independent international, non-governmental, non-profit association (ECSMP, 2016).

³⁴ Covering areas ranging from machine translation, computer-assisted translation, multilingual publishing, speech recognition, dialogue systems and multilingual analytics.

³⁵ The real figure could be higher, as other projects related to human language technologies may have been funded in other research topics.

Guardian, 2014), which was twice as high as the total EC funding to HLT related projects in the last 10 years.

Another initiative that deserves consideration is the program ISA and its successor ISA², that ‘supports the development of digital solutions that enable public administrations, businesses and citizens in Europe to benefit from interoperable cross-border and cross-sector public services’ (European Commission, 2017), and particularly the action of “Overcoming language barriers”, that were intended to open up multilingual tools free of charge that will be easily reusable by SMEs and public administrations. This programme is fully aligned with some of the policies we propose but still the budget is unsubstantial to make a difference, only 160 000 euro in 2016 for the action “Public multilingual knowledge management infrastructure for the Digital Single Market (2016.16)” (European Commission, 2016d).

3.3.3 Multilingualism in national and regional policies

Some EU countries, such as the United Kingdom, Ireland and Spain, have developed interesting strategic plans regarding multilingualism (M. Prys Jones, 2013). Spain, in particular, is a good case of advanced multilingual policies that meet the needs of its linguistically complex population. With nearly 30 % of Spanish citizens being native speakers of a regional language other than Spanish, or Castilian, language policies need to take this diversity into account. Catalan, Basque, Galician and Gascon/Aranese are co-official together with Spanish/Castilian in the regions where they are spoken, while both Aragonese and Asturian are protected within their Autonomous Communities, though they are not recognised as official (Melero et al., 2012).

The article 3 of the Spanish Constitution estates that:

- ‘1. Castilian is the official Spanish language of the State. All Spaniards have the duty to know it and the right to use it.
2. The other Spanish languages shall also be official in their respective Autonomous Communities in accordance with their Statutes.
3. The wealth of the different language modalities of Spain is a cultural heritage which shall be the object of special respect and protection.’

The Spanish Government has specifically supported the development of language technologies with the launch in 2015 of the “Plan to Promote the Language Technologies” (SETSI, 2015). This plan, developed in the frame of a Digital Agenda for Spain (Spanish Government, 2013), aims at boosting the fields of natural language processing and machine translation. The plan targets Spanish/Castilian and the co-official languages, and foresees an investment of 90 million euro in five years. The specific goals of the plan are deployed in 4 axes:

1. Increase the existing linguistic infrastructures. The actions within this axis are:
 - Elaborate and implement a linguistic infrastructure development plan. Infrastructure governance and sustainability.
 - Technical standards for interoperability, license policies and mechanisms of personal data protection.
 - Common tools for resource generation and evaluation.
 - Facilitate the public access to linguistic infrastructure.
2. Boost the national HLT industry and its internationalization, facilitating the development of tools (parsers, semantic annotators, etc.) and linguistic resources (parallel corpora, dictionaries and taxonomies). The concrete actions foreseen are:

- Improvement of the visibility and knowledge transfer between the academic and industrial sector.
 - Support for internationalization and commercialization of the sector.
3. Foster the role of the public sector as a driver for demand, as well as contribute to the standardization and generalization of language resources through the public administration's own activity. Actions designed include:
- Development of platforms for natural language processing and machine translation in the public administrations.
 - Reuse linguistic resources of the public administrations and of public sector information.
4. Develop flagship projects in key sectors such as: health, justice, education, tourism and heritage.

The key elements of this policy focus on:

- Language resources, such as parallel corpora, dictionaries and taxonomies.
- Training professionals.
- Public sector as a driver of demand and as a source of information and data for the development of HLT.

The plan is interdisciplinary and includes various ministries, as well as coordination with the authorities of the Spanish Autonomous Regions, since it also entails their co-official languages.

3.3.4 Multilingualism policies outside EU

We have analysed the multilingual policies of South Africa and India, countries that have multiple official languages and support the development of HLT to identify international best practices that can be applied in the EU (and vice-versa).

3.3.4.1 South Africa

There are eleven official languages in South Africa: Afrikaans, English, Ndebele, Northern Sotho, Sotho, Swazi, Tsonga, Tswana, Venda, Xhosa and Zulu, and the right to use them has been recognised in the country's Bill of Rights.

In 2003, the Government launched a National Language Policy Framework as a tool 'fundamental to the management of our diverse language resources and the achievement of government's goal to promote democracy, justice, equity and national unity' (South Africa Government, 2003). This framework recognised the major role that HLT applications would play in supporting the implementation of the multilingualism policy.

After a decision taken in 2008, in 2010 the National Centre for Human Language Technologies (NCHLT), dependent on the Department of Arts and Culture, was established. The NCHLT funded projects conducted by academic and research institutions.

Later on, in 2011, a Resource Management Agency was created to manage the resources sponsored by the NCHLT. It works as a single depository point for data regarding South African official languages and HLT projects, and represents South Africa internationally. Eventually, it became the hub for language resource management in Africa (LRMA, 2011). The Centre for Text Technology (CTeXt), a research and development centre from the North-West University, is in charge of the Agency. The framework is supported by an Expert Panel (HLTEP) made up of academics and researchers.

A National HLT Network also funds projects, conducts technology audits on language technologies and launches awareness campaigns, as well as promotes networking among educational institutions, private companies and the public sector (South Africa Government, NA).

However, since 2011 there does not appear to have been significant developments in the area.

3.3.4.2 India

There are 22 official languages in India (Hindi, Assamese, Bengali, Bodo, Dogri, Gujarati, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Santali, Sindhi, Tamil, Telugu and Urdu) and 12 scripts. In this context, the Indian Government has placed language technologies in the centre of its ICT policy because it considers that the other 95 % of the Indian population that do not speak English would be excluded from the benefits of the information technology development.

Language technologies are also very relevant for the Indian Government to implement the Right to Information Act, from 2005, that states that: ‘all materials shall be disseminated taking into consideration the cost effectiveness, local language and the most effective method of communication in that local area [...]. A person who desires to obtain any information under this Act, shall make a request in writing or through electronic means in English or Hindi or in the official language of the area in which the application is being made’ (India Government, 2005)

Already in 1991 the Ministry of Electronics and Information Technology initiated the Technology Development for Indian Languages (TDIL) Programme aimed at ‘developing Information Processing Tools and Techniques to facilitate human-machine interaction without language barrier; creating and accessing multilingual knowledge resources; and integrating them to develop innovative user products and services’ (India Government, NA).

In the year 2000, the first major initiative regarding HLT was launched, and the Resource Centres for Indian Language Technology Solutions (RC-ILTS) was established, to promote the development of technological solutions by institutions in concrete languages.

This initiative was followed by the National Roll-Out Plan that continued the work initiated in 1991, and has created and aggregated language software tools and made them available through a web-based Indian Language Data Centre (ILDC), for free. Research on machine translation systems, optical character recognition, on-line handwriting recognition systems, cross-lingual information access and speech processing are supported by the plan. The Government also works in the standardization of language technologies and the inclusion of Indian languages in international standards and supports the local industry with the aim of becoming a multilingual computing hub.

3.3.5 Recommendations of the industry and research in the LT community

3.3.5.1 The Innovation Agenda & Manifesto of LT-Innovate

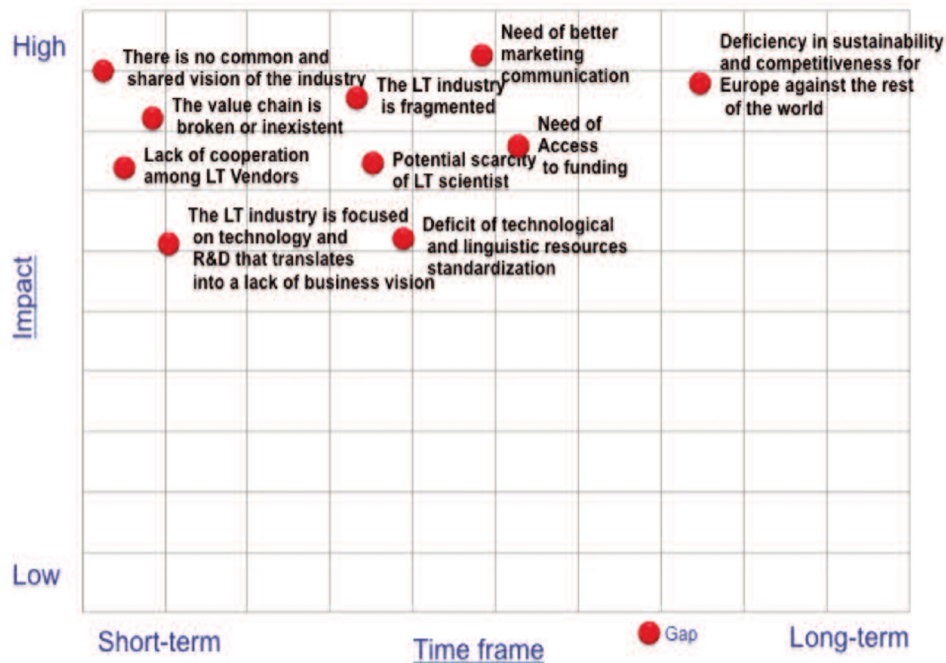
The Innovation Agenda & Manifesto published by the Language Technology Industry Association (LT-Innovate) in 2014 (“Unleashing the Promise of the Language Technology Industry for a Language-neutral Digital Single Market”) indicates that HLT offers many possibilities, but warns that European HLT marketers should focus on people and business applications rather than technology. They alert marketers that sole concentration on technology has created significant barriers to the creation of the right European LT ecosystem.

Some of the effects of this lack of business focus are a broken or inexistent value chain; low understanding of customer needs, lack of coordination between research, LT vendors and integrators or lack of customer orientation. Integrators do not know the value of LT and cannot educate the customer. Moreover, LT is perceived as fragmented industry by integrators, i.e., translation is not connected to speech or to intelligent content. Since the benefits of language technologies are optimised when combined with other technologies (e.g., personal virtual assistants are a combination of speech,

mobile and cloud technologies) and specifically applied, LT vendors should identify precise market needs and determine how integrated LT-based applications can help these segments.

The Manifesto detected ten gaps representing the non-technological challenges in need of attention in order to support the harmonisation and consolidation of the LT Industry. Figure 37 shows these gaps, over which LT Innovate has supported the roadmap outlined in the document, based on the relationship of the level of impact in a short-term/long-term run (time frame).

Figure 37: Non-technological challenges of the LT Industry



Source: Taken from (LT-Innovate, 2014)

Each challenge can be linked to a specific action in the roadmap. Table 5 spells out these actions and links them to each specific challenge.

Table 5: LT challenges and required actions

Gaps / challenges	Roadmap Actions
Lack of common vision and clustering umbrella	Get identified with the LT Industry vision and mission
Fragmentation of the LT Industry	Turn LT Industry into a network of economic advantage
The value chain is broken or non-existent	Enabling the LT value chain for establishing LT as a transversal key technology
Lack of collaboration among LT Vendors	Horizontal integration
Deficit of technological and linguistic resources standardisation	Integrate into a LT ecosystem through collaborative innovation
European LT focus on technology / product	Heading towards a business oriented Pan-European LT Industry
Inadequacy of the proper marketing strategy	Define a go-to-marketing strategy for the third platform and the smart industries
Talent scarcity	Define a new strategy for talent generation and retention
The lack of funded seed	Increase new ventures and entrepreneurs (seed funds)
Lack of European competitiveness	Enable a sustainable and competitive LT Industry: build capabilities through resource sharing, standardization and showcasing

Source: Compiled by the Authors based on (LT-Innovate, 2013)

LT-Innovate's vision embodies the following strategic features of Europe's digital future:

- Mastering human language is the next big opportunity for ICT; LT is a key technology of the future.
- LT is a maturing industry; there is strong global demand for LT in mainstream products and services.
- European companies could command a larger share of the global LT industry, but need to overcome fragmentation and market barriers.
- There is no lack of innovation in Europe; the main stumbling block is that European SMEs do not grow beyond their national or regional linguistic islands to address European and global markets.
- Europe needs a shared language infrastructure to preserve its languages, give people and organisations access to the Single Digital Market using their own languages, and underpin innovation in the LT industry itself.

Figure 38 presents graphically how the three LT main segments or technologies (intelligent content, translation and speech) can have an impact on the different horizontal market segments.

Figure 38: Impact of LT on horizontal segments

	Software & Services for the Management of								
	Business Intelligence	Content	Customer Experience	Sales	Marketing	Human Resources	Supply Chain	Legal/Risk	Finance
Intelligent Content Technology	*	*	*	*	*	*	*	*	*
Translation Technology	*	*	*	*	*	*	*	*	*
Speech Technology	*	*	*	*	*	*			

Source: Taken from (LT-Innovate, 2013)

3.3.5.2 META-NET Strategic Research Agenda for Multilingual Europe 2020

More than two years in the making, the final version of the META-NET “Strategic Research Agenda for Multilingual Europe 2020” (SRA) was published on December 01, 2012 (Rehm & Uszkoreit, 2012). The META-NET SRA is the result of a discussion between hundreds of experts from research and industry. The main purpose of the SRA is to raise awareness in the field of language technology in Europe and attract the attention of and inform politicians and policy makers of the regional, national and international level in their decisions, especially with regard to Horizon 2020 and Connecting Europe Facility (CEF).

The META-NET SRA recommends three priority research themes: (1) Translingual Cloud; (2) Social Intelligence and e-Participation; and (3) Socially Aware Interactive Assistants. These three themes are complemented by two themes that focus upon (4) core technologies and resources for Europe’s languages and (5) building the European service platform for language technologies.

3.3.5.3 Strategic Research and Innovation Agenda of the ‘Cracking the Language Barrier’ federation

With the goal of addressing the challenge of language-induced market fragmentation in mind, the ‘Cracking the Language Barrier’ federation³⁶, which consists of 11 organisations and more than 20 projects, has prepared a Strategic Research and Innovation Agenda (SRIA), with the title “Language as a Data Type and Key Challenge for Big Data”, a first version of which was presented at META-FORUM

³⁶ <http://www.cracking-the-language-barrier.eu/>

2016 in Lisbon, Portugal, with a final version expected in 2017. This document recommends a three-year Multilingual Value (MLV) Programme capable of enabling the Multilingual Digital Single Market.

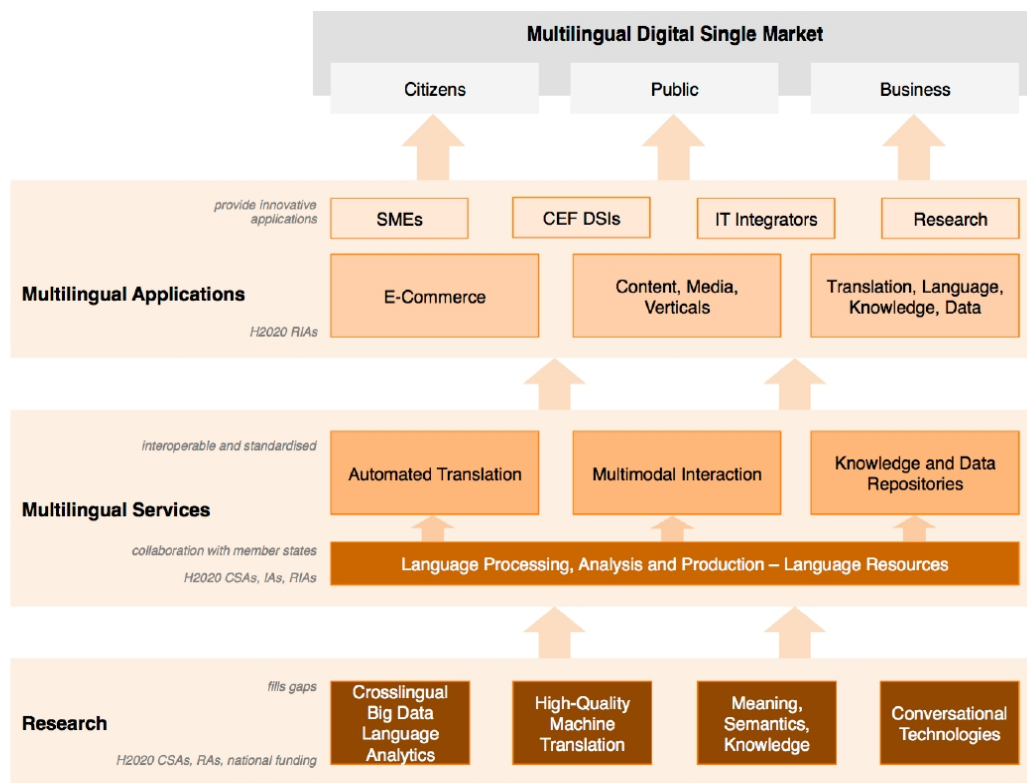
The MLV Programme consists of three application areas that relate to the three main pillars of the Multilingual Digital Single Market:

1. Multilingual E-Commerce: provides multilingual and cross-lingual technologies around search, customer-relationship management, helpdesks, processes, workflows, product catalogues and descriptions etc.
2. Multilingual Content and Media: assembles cross-lingual technologies for content analytics, curation and generation including authoring support, multimodals and social media.
3. Translation, Language, Knowledge, Data: provides multilingual applications that connect big data technologies and language as well as knowledge technologies including machine translation (written, spoken, automatic/human), text mining, business intelligence, sentiment analysis, domain-specific approaches and semantic annotation.

These applications are driven by several multilingual services, which are, in turn, fostered and further improved through research. There is also the plan to intensify work on basic technologies so that all relevant languages are covered. In addition, horizontal topics need to be addressed, e.g., standardisation, interoperability, and policy aspects.

The aim of the MLV Programme is not only 'unlock the multilingual Digital Single Market through a set of platforms, services and solutions that support all businesses and citizens, but 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.'

The MLV Programme also aims at reducing the threat of digital extinction for many European languages. This Programme recommends that 'Europe actively makes an effort to compete in the global landscape for research and development in language technology since we cannot expect third parties from other continents to solve our translation and knowledge management problems in a way that suits our specific communicative, societal and cultural needs.'

Figure 39: The Multilingual Value Programme

Source: Taken from (SRIA, 2016)

The SRIA 2016 recommends starting with a small set of clearly defined services needed by most of the applications. This initial set of seed services would then scale organically into one or more bigger platforms, ideally a shared European HLT platform. Current resource-exchange infrastructures, such as for example META-SHARE, will play an important role in the design of such a platform, which initially will need 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, it will need to be established by one or more consortia with strong commercial partners and also be operated by these consortia or commercial contractors.

Thanks to the HLT platform approach:

- HLT 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. Providers of other services that can be enabled or enhanced by HLT will use the platform for testing the needed LT functionalities and for integrating them into their own solutions.
- HLT consumers will receive integrated services without having to install, combine, support and maintain any software, having easy access to specialised solutions even if they do not use them regularly. Corporate users will enjoy the benefits of HLT early and at no (or reasonable) cost through a large variety of generic and specialised services offered through a small number of sources. 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.

The SRIA 2016 presents a detailed roadmap for the deployment of the MLV Programme. The estimated costs for its implementation is in the range of 175-200 million euro for the first three years and phases (2018, 2019, 2020), including industry contribution (ca. 20 %). After 2020, a substantial reduction of funding is expected since the multilingual services should become self-sustained and profitable, while the multilingual applications will include a commercial component right from the start. At the same time, funding for research would need to be maintained in order to secure a leading position for Europe in the HLT field.

These costs will be much compensated by the expected benefits of the programme. The EC predicts that the transition to the integrated Digital Single Market will deliver up to 400 billion euro in economic growth by 2020. However, this ambitious goal 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.

The SRIA concludes that ‘only if Europe accepts the multilingual challenge and decides to design and 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’, and that ‘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.’ They further expect that if the strategic programme is successful, Europe could offer the developed solutions to other multilingual societies, for example, to adapt and to export certain parts of the MLV Programme to India or South Africa.

The suggestion is that the EC supports the MLV Programme especially through dedicated activities in upcoming Horizon 2020 calls (2018–2020) and through Connecting Europe Facility (CEF). The European HLT industry (mostly SMEs) will then participate in the practical application areas, where highly innovative activities with a major commercial impact are needed. 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.

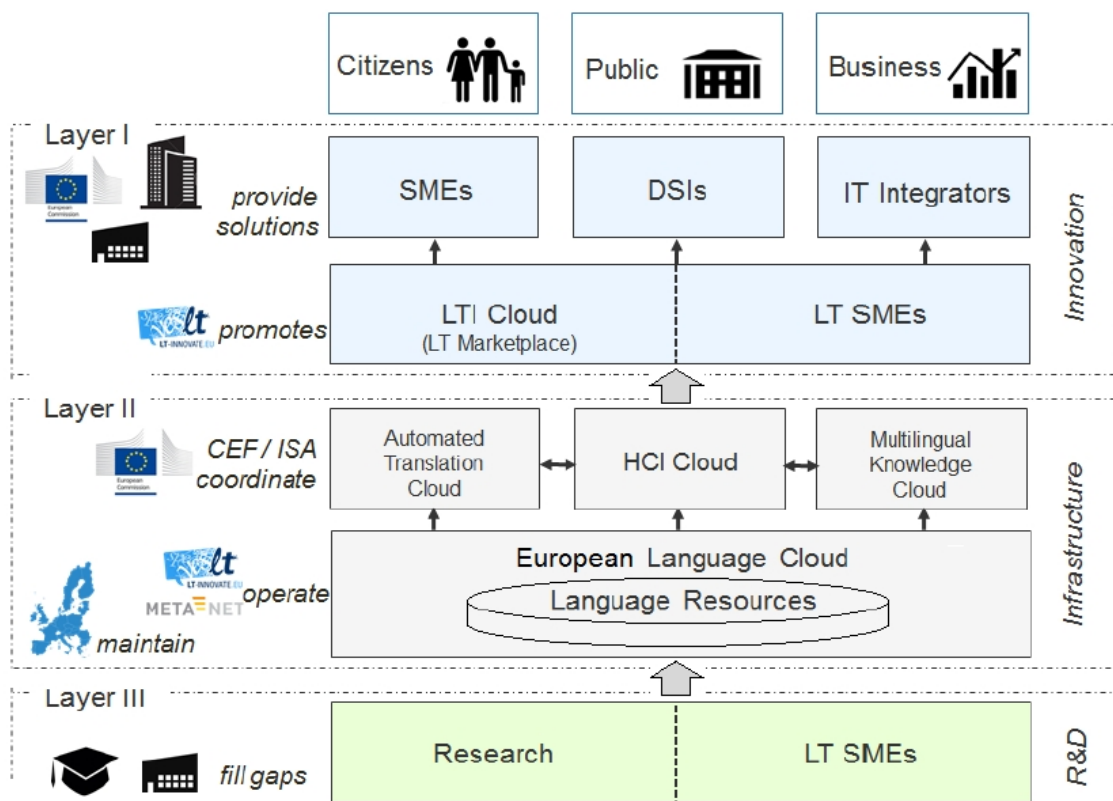
3.3.5.4 European platform for multilingual DSM (LT-Innovate, 2015)

The deployment of cloud computing solutions in the modality of Infrastructure as a Service (IaaS) is going to facilitate the development of HLT services, given the benefits that they provide:

- Scalability: the computing resources for training NMT systems can be adapted, increasing or reducing it, almost in real time.
- Transforming CAPEX into OPEX: IaaS solutions eliminate the initial huge cost of the IT infrastructures required to support NMT, allowing a pay-as-you-go model, more suitable for SMEs that can develop their own LT services.
- Accessibility: the companies can access the computing resources from any location.
- Security and service continuity: the cloud computing providers ensure the physical and logical security and the continuity of the service under any circumstances.

At European level, the industry association LT-Innovate has suggested the creation of a European Platform for Multilingual DSM (LT-Innovate, 2015), which would encompass a set of clouds. One of them, the Automated Translation Cloud, would be devoted to providing high-quality machine translation services as well as linguistic resources as needed.

Figure 40 shows the diagram of this European Platform.

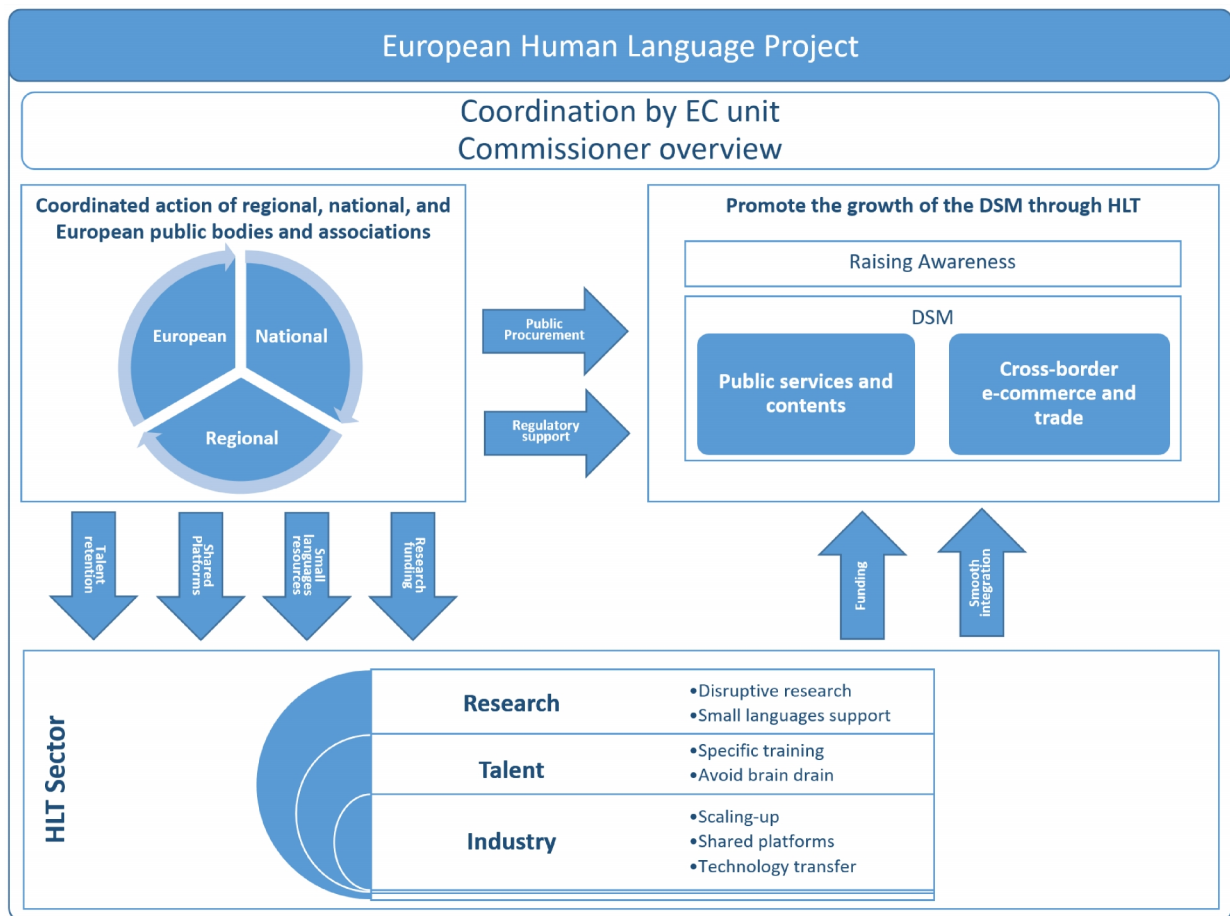
Figure 40: Diagram of the European Platform for the Multilingual Digital Single Market

Source: Taken from (LT-Innovate, 2015)

4 Policy options

Multilingualism in Europe is a complex topic involving many stakeholders with intertwined interests in different countries. Therefore, no single policy can tackle the problems described in the previous chapter. On the contrary, to truly seize the opportunities of the multilingual Europe we believe that a joint and coordinated action at the European, national and regional levels involving stakeholders from the public sector, civil society, research institutions and industry is required. To do so, the main recommendation is to launch of a multidisciplinary European Human Language Project including multiple actions as shown in Figure 41.

Figure 41: The European Human Language Project



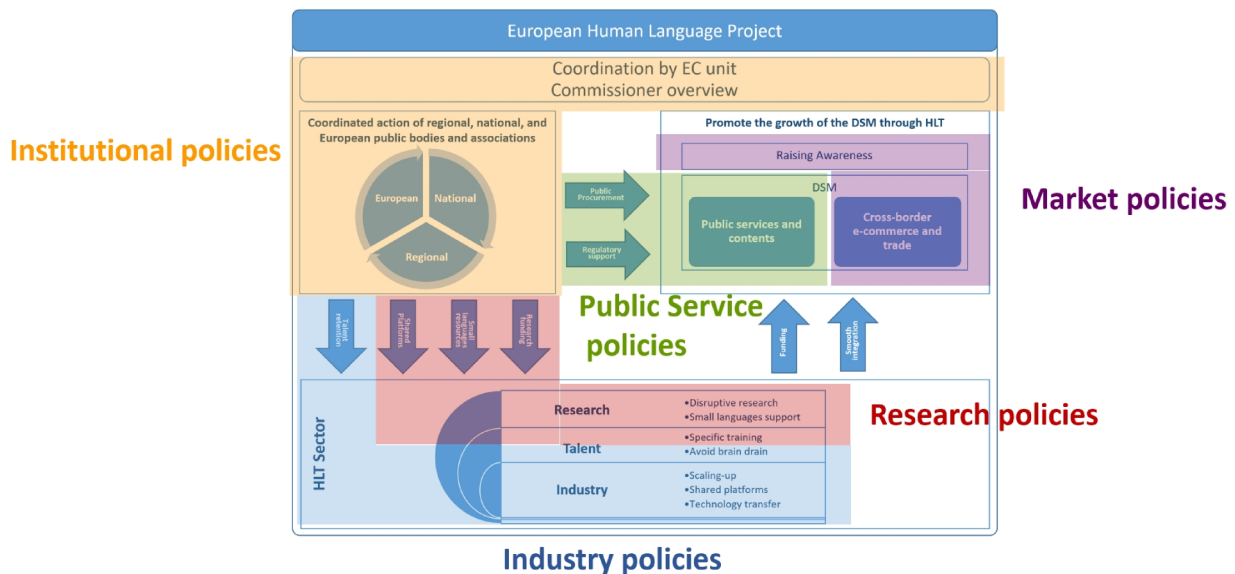
Source: Compiled by the authors

To fully achieve the potential of HLT in the multilingual Europe a coordinated joint effort is required. We suggest that the EC should be in charge of coordinating the policies at the different levels. Not only European institutions, but also national and regional governments should be responsible of creating resources for their languages. Research in Europe should focus on fostering the new deep learning paradigm in HLT, while at the same time providing support for smaller European languages through technology transfer. Talent scarcity and drain brain should be transformed into talent creation and brain gain. Coordination between research and industry should be provided in a seamless, open and effective way through existing European platforms. The public sector should provide their contents and services for all European languages while promoting the growth of the HLT market through public procurement of innovative technology. Mechanisms to facilitate the scaling-up of European innovative HLT

companies should be established. Eventually, policies supporting firms across Europe that sell cross-border by providing their contents, products, and services in the different European languages should be enacted to create a fully integrated DSM.

To achieve these goals, a set of policy options, which are structured into five groups, is proposed and assessed: institutional policies, research policies, industry policies, market policies and public services policies, as outlined in Figure 42.

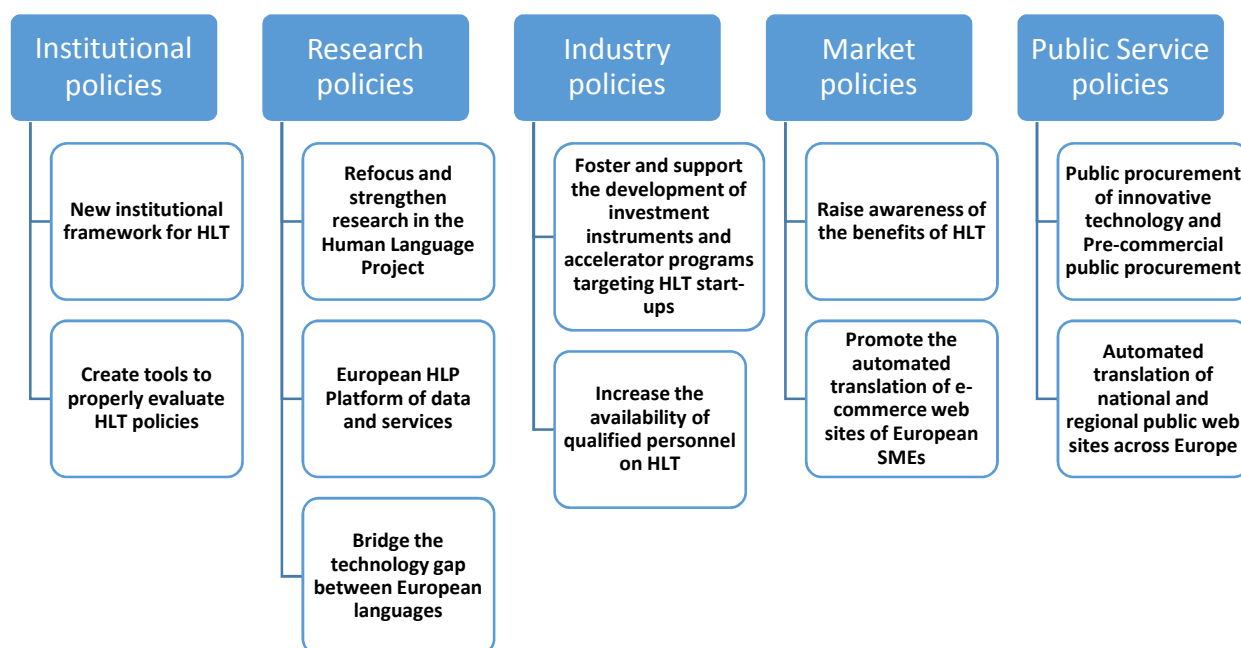
Figure 42: The European Human Language Project and proposed policies



Source: Compiled by the authors

Figure 43 shows a more detailed picture of the policies. Institutional policies involve initiatives to evolve current institutional framework to draw upon emerging technology trends to better fit the challenges of the multilingual Europe while properly assessing the results. Research policies focus on moving Europe to develop the next generation of Language Technologies. Research policies also aim at integrating research and industry, providing Europe with the tools to share resources to effectively compete with other markets, while at the same time contributing to the equality of all European citizens in their everyday digital experience, regardless of their language. Industry policies foster the creation and growth of competitive European firms while increasing the availability of highly qualified workers. Market policies seek to improve the HLT sector in Europe by raising awareness among European stakeholders of the relevance of these technologies to further increase the demand of services. There is a specific policy targeting small web merchants so they can benefit from accessing a much bigger market by translating their web shops by using HLT. Public service policies intend to create multilingual public services in European, national, regional and local administrations while contributing to the increase the innovative HLT sector by using public procurement tools.

We have followed an approach where the different policy options reinforce each other instead of being mutually exclusive.

Figure 43: Proposed HLT policies

Source: Compiled by the authors

The structure of the chapter is as follows:

1. Based on the findings of the analysis, the main challenges facing Europe are described.
2. Based on these challenges, a matrix of assessment criteria is constructed.
3. Different policy options are described and assessed using the criteria described in chapter 4.2.

4.1 Challenges

Based on the results of the analysis, we have classified the main challenges multilingual Europe is facing into four groups: institutional, social, economic and sector challenges.

4.1.1 Institutional challenges

Our analysis suggests that HLT are not properly represented in the European policy agenda and, therefore, there is a low involvement on the part of the European institutions on finding effective solutions to the issue of language barriers in the DSM by using innovative language technologies. There is also a lack of coordination between initiatives at the European, national and regional levels.

4.1.2 Social challenges

All European citizens need to be enabled and empowered to communicate and to operate in their mother tongues, online as well as offline. European policy makers have to future-proof our languages so that the coming generations will be able to use them online and offline. However, in the current

situation there is an unfair language gap and many European languages are in danger of digital extinction.

Unfair language gap

We have seen that language barriers are fostering and widening an unfair language gap among citizens, and among linguistic communities, which is particularly challenging in the digital era. Less educated and older citizens are left behind while communities with fewer speakers will be more deeply affected by the language gap. Language equality must be seen and treated with the vision that multilingualism is at the very heart of the European idea. The European Union has 24 official languages; all have and enjoy the same status. In addition, Europe has dozens of regional and smaller languages as well as languages of immigrants and trade partners. We should be able to empower all citizens to use their mother tongues, online as well as offline.

Smaller languages in danger

Not only official languages should be considered, but also smaller languages, sign languages, migrant languages and languages of trade partners. There are dramatic differences among Europe's languages in terms of both the maturity of the research and the state of readiness with respect to LT solutions. The smaller the language, the fewer, and of lesser quality, technologies there are available for that language. Even EU official languages with number of speakers below 10 million suffer from a weak LT support.

There is an insufficient access to language resources and to public or open data in most European languages. In many cases, even if some resources are there (language corpora, etc.), public institutions are not sufficiently aware of the importance of making them available to researchers and developers.

4.1.3 Economic challenges

Making Europe's Digital Single Market multilingual thanks to language technologies is one specific challenge field, one in which LT can be applied to broaden the reach for markets with thousands of ecommerce services and online shops in Europe. At the same time, there are additional missions and opportunities as well as other sectors in which LT can be used and deployed for multiple positive effects. However, the availability of public services and contents in the DSM is very low and the DSM e-commerce is strongly fragmented in almost isolated markets.

Low availability of multilingual public and private services

There is a low percentage of firms providing their on-line service in multiple European languages. Moreover, the web pages of national and regional public institutions usually provide their services in very few foreign languages. This might suggest low awareness about the status of the technology and its potential benefits.

Profound effect on the Digital Single Market

Language barriers are likely to have a profound effect on the creation of a truly integrated and fair Digital Single Market. Language barriers affect e-commerce, leaving those citizens that are unable of speak the major languages out of the benefits of a stronger and more effective market, while hindering the mobility of workers and the possibility of accessing public services in other countries. Furthermore, language barriers are market barriers. Customers who cannot understand the language of an e-commerce website that could potentially be of interest will not purchase anything from said website. If the Digital Single Market is not inherently multilingual, there will be more than 20 isolated markets fragmented by their respective languages.

4.1.4 Sector challenges

One of the key growth topics of the next years will be artificial intelligence. Almost on a weekly basis, important breakthroughs are being reported, massive investments are being made, especially in the USA and in Asia. Meanwhile the EU is investing in language technologies, but only on a rather small scale. As a consequence, many researchers move to non-European companies. Regarding the industry, the European HLT sector is fragmented in small companies with broken or inexistent value chains.

Basic research is still needed

Language technology is still far from being perfect. According to the META-NET White Paper Series (Rehm et al., 2014, 2016), no single language (not even English) can be considered to have an “excellent” technological support. LT as such still need lots of research in order to make significant improvements to lead them to their full potential.

Lack of talent

There is talent scarcity combined with the human capital flight problem that hinders the development of the industry. Many high potentials and researchers are going to the USA, where they get very high salaries. We need to find incentives to make them stay, both by creating an environment of talent as well as fostering a promising start-up ecosystem.

Strong fragmentation of the European industry

Although there is no lack of innovation in Europe, there is a strong fragmentation of the LT industry. The main stumbling block is that European SMEs do not grow beyond their national or regional linguistic islands to address European and global markets. Foreign multinationals, mainly American, are drawing upon European talent and technology to create their offerings through the acquisition of innovative small and medium European companies.

Lack of European language technology ecosystem

There are several challenges for the European HLT ecosystem: (1) lack of coordination and missing links between research, LT vendors, integrators and customers; (2) insufficient demand from public sector, especially compared to the USA where defence, intelligence and other government sectors are the main drivers of LT industry; (3) market disruption by global players like Google Translate, who subsidize LT services from their dominating position in other sectors; (4) insufficient market in smaller countries to justify investments in LTs for smaller languages that tend to be expensive for lack of previous research and resources; and (5) low understanding of customer needs while, at the same time, integrators do not know the value of LT and cannot educate the customer.

4.2 Assessment criteria

The criteria are selected based on the problems described in the previous chapter:

1. **Cost and benefits:** whether the benefits fostered by the policy clearly surpass the costs and whether the policy is affordable for both public and private agents.
2. **Political feasibility:** whether the policy is likely to be backed by political support. This includes balancing the mandate-term vision of politicians with long-term policy vision.
3. **Feasibility in the EU context:** whether the policy can be easily implemented in the current European framework.
4. **Effectiveness:** whether the policy is expected to have a feasible cost per target achieved.
5. **Sustainability:** whether the policy can be kept over a sustained period of time.

6. **Risks and uncertainties:** whether the policy is likely to achieve its goals or if there is high uncertainty or risks associated to the policy

7. **Coherence with EU objectives:** whether the policy fits with EU objectives, such as constructing a European identity, preserving cultural values, etc.

8. **Potential ethical, social, environmental and regulatory impacts:** whether it is expected that the policy has unexpected consequences regarding some aspects subject to general interest or constituents' scrutiny.

9. **Innovation:** whether the policy contributes to increasing innovation in Europe.

10. **Single market:** whether the policy contributes to creating a single market in Europe while reducing disparities among the countries.

11. **Economic growth and job creation:** whether the policy contributes to overall economic growth, particularly improving the labour market.

12. **Tackling inequalities:** whether the policy narrows the language divide.

The different policy options are assessed against each of these criteria by using a three-level scale: low if the policy does not adequately meet the objectives, high if the policy is likely to meet the objective, and medium, someplace in between. A fourth value of "uncertain" is applied if the effect of the policy remains unclear.

4.3 Institutional policies

4.3.1 Reinforce the role of HLT within the institutional framework of multilingualism related bodies

Our first recommendation is to reorganize HLT activities within the EC by having a unit at the highest level possible to take care of "Multilingualism and Language Technologies" to increase efficiency, coordinate the different policies and to allow for a more holistic approach. There are several options, such as combining all the current Commission's activities/units in support of HLT or setting up an additional new unit in the EC to coordinate with all existing units that are related to the topic. The Commission should take into consideration the decision on how to implement this proposal if it is agreed upon.

This new organization should provide services and advice to all European public and private bodies, including the EC itself. This unit will be in charge of the definition of the policies regarding HLT in Europe. This unit should also coordinate the efforts of national and regional organisations as well as funding agencies. Moreover this unit should be in charge of carefully assessing the results of the different European, national and regional policies related to the topic as proposed in the following policy option.

In addition, we suggest attaching the "Multilingualism and Language Technology" responsibility to a Commissioner. Currently no Commissioner has this topic in their portfolio and this needs to change. "Multilingualism and Language Technologies" or "Multilingualism through Language Technologies" is so important to Europe and its structure that it must be part of a Commissioner's tasks and responsibilities.

Evidence and assessment

In the European Commission there is a Unit, called "Learning, Multilingualism, and Accessibility (Unit G.3)", within the Directorate G of the DG Connect, which is in charge, among other things, of supporting 'policy, research, innovation and deployment of learning technologies and key enabling digital language technologies and services to allow all European consumers and businesses to fully benefit

from the Digital Single Market' (European Commission, 2016i). Because of the relevant role that the HLT are likely to play in the integration of the EU, these technologies by themselves should have a specific unit at the highest level, independent of the learning and accessibility aspects.

Moreover, in the DG Translation there are specific units of "Language applications" and "Machine translation" (European Commission, 2016b) that could further support national and regional governments, and, in the DG Informatics, there is the "ISA unit" (DIGIT.B6), which is in charge of the ISA² programme, also related to overcoming language barriers (European Commission, 2017).

Regional, national and international organisations as well as funding agencies should team up to support shared programs to develop at least basic resources and technologies for all European languages. Such a large-scale effort is needed to reach the ambitious goal of providing support for all European languages, for example, through high-quality machine translation and the creation of high-quality linguistic resources for all languages. This depends on many stakeholders, such as politics, research, business and society all uniting their efforts.

Table 6: Assessment matrix for the "Reinforce the role of HLT within the institutional framework of multilingualism related bodies" policy

Criteria	Adequacy	Argument
Costs and benefits	High	The cost of integrating existing units is expected to be low while the benefits of a higher coordination can be very high in the medium term.
Political feasibility	High	This is an organizational policy within the EC and it is not expected to be subject to political controversy.
Feasibility in the European context	Medium	While the creation of the unit itself is not likely to be problematic, the coordination and assessment of national policies may be more controversial.
Effectiveness	High	The effect of having a unique contact point for these policies is likely to be much more effective than the current situation.
Sustainability	High	The unit is likely to easily evolve over time to adapt to new requirements and specificities of the environment.
Risks and uncertainties	Medium	There is a risk that the unit will not be able to effectively coordinate the actions particularly if the unit continues focusing on internal requirements instead of having a wider approach.
Coherence with EU objectives	High	One of the EU objectives is to foster a truly integrated and fair Union. Effectively tackling language barriers is clearly one of the tools to achieve that target.
Potential ethical, social, environmental and regulatory impacts	Medium	The unit can promote legislation to foster the use of HLT and therefore can have a regulatory impact.
Innovation	Medium	The unit can positively contribute to improve research and innovation of HLT in Europe, although it should be done by coordinating efforts and policies of third parties that will not be under its control.

Single Market	High	The unit can significantly contribute to improving the DSM by tackling language barriers.
Economic growth and job creation	Medium	The unit can contribute to a more competitive market. However, an extensive use of innovative technologies can also contribute to job destruction among those workers who can be replaced by technology.
Tackling inequalities	High	By tackling language barriers, it will contribute to shrink the language divide.

4.3.2 Create tools to properly evaluate HLT policies

In order to properly assess the results of the policies it is very important to create specific surveys and datasets that could facilitate the assessment of HLT policies. This should be done when defining the new policies, and they should collect information before and after the policies are enacted. Moreover it will be particularly interesting to create open³⁷ and homogeneous European longitudinal datasets including extensive information about language barriers, its economic and social impact, and the effect of HLT on overcoming these barriers and on the expected outcomes. It will facilitate the assessment of the effects over time.

This policy is particularly relevant in Europe, which can act as a natural policy lab where different policies and approaches can be tested at the national and regional levels, providing a valuable experience for other countries to define their policies based on previous experience. Take the “Plan to Promote the Language Technologies” (SETSI, 2015), which is being developed by the Spanish Government. This plan should provide valuable insights for other European countries when defining policies to support their own languages; however, without carefully assessing its results the opportunity will be lost.

Evidence and assessment

In defining new policies it is very important to learn by experience. One of the main challenges affecting innovative policies is the lack of reliable information about the effects of the public intervention on the outcomes. Without that information it is difficult to carry out further research to get on whether or not to support the decision of maintaining, reinforcing or cancelling current policies and whether or not to design new policies. In fact, when undertaking this analysis, it has been difficult to find specific data about HLT, language barriers and policies related to the topic. We have been obliged to base our analysis on other datasets containing less reliable, indirect information.

Researchers and analysts could draw upon this dataset to assess the benefits of the different HLT policies. Moreover, this dataset should include information about short and medium-socio-economic outcomes. The EU can promote the creation of this knowledge at the European level that can further be used to advise national and regional governments about their specific policies.

³⁷ Excluding sensible data or data that cannot be anonymised.

Table 7: Assessment matrix for the "Create tools to properly evaluate HLT policies" policy

Criteria	Adequacy	Argument
Costs and benefits	Medium	Collecting the required information can be costly while the effects are likely to arise slowly. This can hinder the deployment of this policy.
Political feasibility	Medium	Short-term interests could hinder this policy because the results are likely to arise in the medium and long-term. Moreover, politicians might not always be interested on transparency about the effects of the policies they enact.
Feasibility in the European context	Medium	Creating this knowledge can be costly and challenging in the European environment, which includes 28 countries with specific interests.
Effectiveness	Medium	Although it can have a profound effect on the effectiveness of the policies in the long-term, the results in the short term are likely to be low. The good news is that Europe is a natural policy lab where different policies can be easily tested and assessed.
Sustainability	Low	The policy should be sustained over time, particularly if it means creating longitudinal datasets.
Risks and uncertainties	Medium	Providing quantitative evidence can be dangerous if the analysis and conclusions are flawed (Gorard, 2014) or are opportunistically used (Group, 2014).
Coherence with EU objectives	High	The EU is committed to define policies based on evidence. Moreover, current budget pressure is forcing policy-makers to define more effective policies.
Potential ethical, social, environmental and regulatory impacts	Medium	This policy might benefit disadvantaged population and therefore is likely to have a positive effect on ethical and social considerations. It is not expected any change on environmental considerations.
Innovation	Low	Although defining more accurate policies will likely promote innovation, big improvements are not expected in this area.
Single Market	High	By defining the policies based on evidence, it is expected that language barriers will be more effectively overcome.
Economic growth and job creation	Medium	Although defining more accurate policies will likely promote economic growth in the medium term, big improvements are not expected in this area.
Tackling inequalities	High	The results of policies to tackle language barriers will be substantially improved by drawing upon knowing the effect of the policies on disadvantaged population.

4.4 Research policies

The main goal of the suggested research policies is to initiate a large-scale research and development and innovation (R&D&I) funding programme in order to provide robust, precise and attainable technology solutions for deep natural language understanding by 2030. If set up correctly, such a long-term, large-scale programme would address all aforementioned issues and missions and, at the same time, propel Europe into the top position with regard to the highly crucial sector of language technologies. This push is needed by the European LT research community and also by the European industry that provides LT to commercial clients. This push is needed to future proof our languages and to save our languages from digital extinction. This push would contribute to multilingual digital public services and also to the multilingual Digital Single Market. This push would play a huge role in empowering all European citizens to use their mother tongues, online as well as offline.

We also intend to provide the European LT community with the ability to compete with other markets by effectively sharing resources and services, and to contribute to the equality of all European citizens, regardless of their language, in their everyday digital experience. In order to accomplish this, we recommend a common European LT platform connecting LT-providers and LT-consumers, and a concerted effort to foster research and technology transfer among languages and scientific communities across Europe, with the purpose of reducing the language technology gap.

4.4.1 Refocus and strengthen research in LT through a Human Language Project

The first and probably most crucial research policy that we recommend in this report is to initiate the Human Language Project (HLP) as a large-scale, long-term R&D&I funding programme, in which basic research, applied research, development, innovation and commercialisation work closely together in order to develop technologies for deep natural language understanding by the year 2030 (Rehm, 2017)³⁸. The research areas and fields needed to contribute to this goal are: computational linguistics, linguistics, artificial intelligence, language technology, computer science, cognitive science and several others. An important aspect of the HLP is not only to support and boost, but also to coordinate all European R&D&I activities in a systematic and targeted way by including the European Parliament, European Commission and the Member States. Important outcomes of the programme will be, not only ground-breaking methods, paradigms and approaches, but also technologies and products, as well as a fostered economy.

The key goal and challenge of the research activities of the Human Language Project needs to be nothing less than deep natural language understanding (including language generation), i.e., the main R&D goal of the whole field. Current LT approaches are still rather shallow, many research and technology breakthroughs are needed in order to move closer to the goal of full natural language understanding. Nevertheless, recent breakthroughs in artificial intelligence and a fresh look at recent results in linguistics can bring about the needed technology shift for the next generation of LT. With the HLP, Europe would be in an excellent position to achieve research and technology leadership in this important field of ICT. With regard to research, key ingredients of the HLP should be artificial intelligence, language technology, machine learning and knowledge technology. Along with these, HLP includes certain side topics such as, among others, cognition, perception, vision, cross-modal and cross-platform. In terms of languages, all official and many non-official European languages need to be addressed, plus several others being currently spoken by European migrants. Technological goals

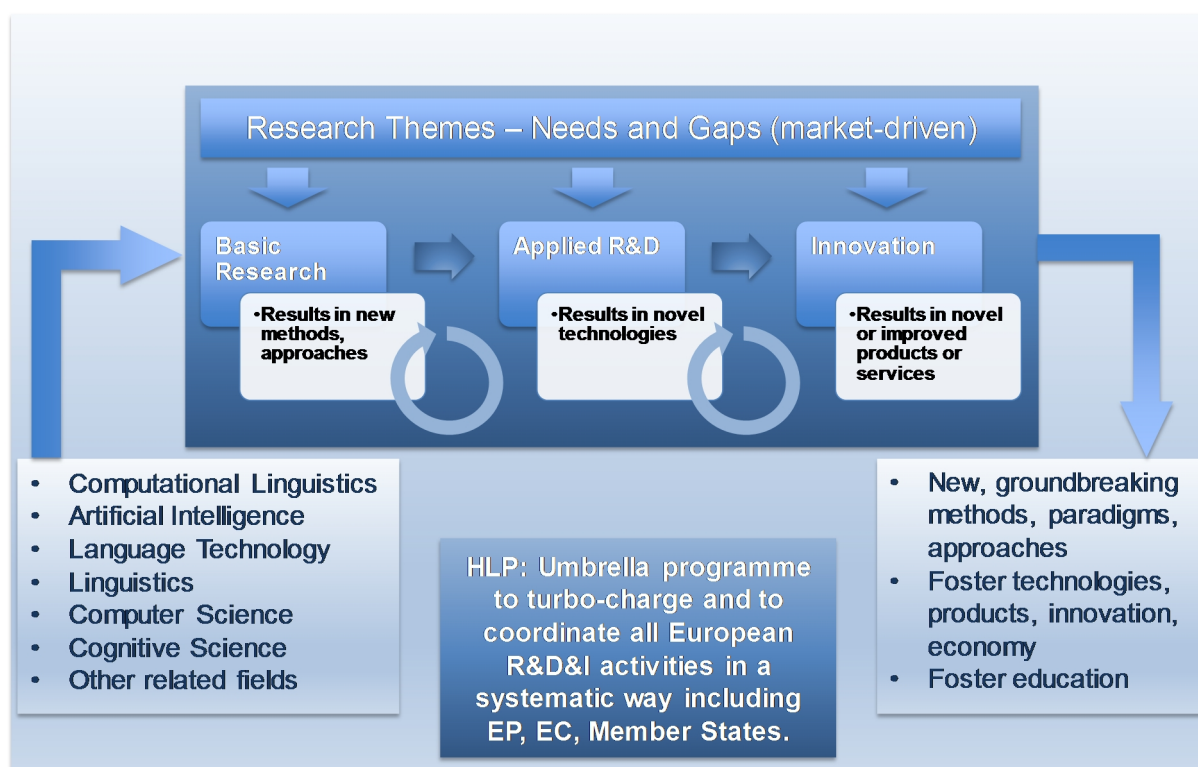
³⁸ As described by Georg Rehm, co-author and co-editor of the META-NET SRA (2013) and the Multilingual DSM SRIA (2016) in his presentation “Human Language Technologies in a Multilingual Europe”, STOA Workshop on Language Technologies, European Parliament, 10 January 2017.

should be broad coverage, high quality and high precision. The technologies that need to be developed should be able to work across modalities (text, text types, speech, image, video etc.), across platforms (messaging, telephony, social, mobile, internet of things etc.), and across cultures (knowledge, customs, formalities, humour, emotion, subjectivity, biases, opinions, filter bubble, etc.).

It is worth mentioning that the multilingual value programme, as suggested by the strategic research and innovation agenda for the multilingual Digital Single Market, can be conceptualised as part of the Human Language Project. The MLV programme addresses an important part of the European language challenge but it is, by no means, the only part. The challenge itself is bigger and, hence, the solution needs to address several additional dimensions. The MLV programme identifies four priority research themes that should support and further improve the LT-based services and applications that aim at setting up the multilingual Digital Single Market. Those research themes are: high-quality machine translation; cross-lingual big data analytics; conversational technologies; and meaning, semantics and knowledge technologies aimed at exploiting the data value chain.

In brief, our suggestion is to set up, under the umbrella of the HLP, a coordinated initiative both on the European (EC/EU), and national and regional levels (Member States, associated countries, regions), including research centres and small, medium and large enterprises that work on or with LT and other stakeholders, especially user companies. A summary of the research strategy of the HLP is shown in Figure 44.

Figure 44: HLP Research strategy



Source: Taken from (Rehm, 2017)

In terms of funding sources, a mix can be foreseen between Horizon 2020 (Work Programme 2018-2020), the next Framework Programme (until 2021) as well as national and regional funding sources. The setup must include the whole chain of research, i.e., basic research, applied research and development as well as innovation and commercialisation. The HLP must have a runtime of at least ten years in order to be effective and produce really ground-breaking results. At the same time, a significant policy change is needed in the EU and in the Member States towards "LT-enabled Multilingualism" and the use of

sophisticated LT for several typical communication scenarios, both online and offline. This would produce the required pull effects.

Evidence and assessment

The European Union is currently investing in language technologies, but only on a rather small scale. Under the Connecting Europe Facility (CEF) programme, a service for automated translation is being developed, which is already used to make digital public services multilingual. The corresponding work is carried out by the EC, supported by a small set of service contracts, among others, for collecting datasets and language resources. Within Horizon 2020, smaller budgets are available in the big data part of the ICT work programme, mainly for innovation projects but also for some research and innovation projects, in which language technologies are used to build cross-lingual data value chains. These current initiatives are too small, too focused and too unbalanced and they are concentrated on innovation and technology deployment. When too much focus is given to driving innovation, there is always a danger of losing touch with basic research and potentially paradigm-shifting developments. Furthermore, it is becoming increasingly difficult to kick-start new, potentially paradigm-shifting research initiatives. This is why a broader, coordinated, concerted and consolidated push in basic research, applied research and development and innovation is needed.

There is also a lack of coordination between European research and the market of applications and services. The MLV Programme proposed by the SRIA 2016 aims to address this important challenge by recommending a concerted European effort of administration, research and industry to compete in the global landscape of language technology.

Table 8: Assessment matrix for the "Refocus and strengthen research in HLT" policy

Criteria	Adequacy	Argument
Costs and benefits	High	Costs will be much compensated by the expected benefits. Already during its early years, the HLP would generate benefits and have positive effects with regard to multiple areas of life: it would significantly stimulate the European economy (by enabling the Multilingual Digital Single Market), it would create attractive jobs for high potentials, it would provide ground-breaking novel input for the education of our young researchers, it would produce a serious boost for research, it would foster innovation and new companies and it would help create a truly multilingual Europe. We expect the economic growth by 2020 due to the integrated Digital Single Market to be much higher than the predicted 400 billion euro since we would have successfully enabled many European SMEs to participate in the multilingual DSM, substantially multiplying their reach.
Political feasibility	High	Horizon 2020 calls (2018-2020), the next framework programme, and CEF are the perfect EU funding instruments for this policy. 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.
Feasibility in the European context	High	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 to Europe, the EU itself would be the perfect reference user of such technologies, setting an example for national or regional governments.
Effectiveness	High	A key factor for the success of the programme endorsed by this policy is that it has been consensually elaborated by the main actors in the European Language

		Technology community – including research, development, innovation and other relevant stakeholders.
Sustainability	High	Funding for research would need to be maintained in order to secure a leading position for Europe in the HLT field. For instance, the SRIA 2016 presents a detailed roadmap for the deployment of the MLV Programme. The estimated costs for its implementation is in the range of 175-200 million euro for the first three years and phases (2018, 2019, 2020), including industry contribution (ca. 20 %).
Risks and uncertainties	Low	In order to lower uncertainties, it is recommended to start with a small set of clearly defined services needed by most of the applications. This initial set of seed services would then scale organically into one or more bigger platforms. This strategy will reduce risks in its deployment.
Coherence with EU objectives	High	The programme is based on existing EU funding instruments whose goals are perfectly aligned with the programme's objectives.
Potential ethical, social, environmental and regulatory impacts	Medium	Social impact will undoubtedly be positive. We are not aware of any regulatory conflict with the spirit of this policy. Ethical and environmental probably do not apply.
Innovation	High	The integrated LT ecosystem resulting from the application of this programme will undoubtedly promote and accelerate collaborative innovation.
Single Market	High	The aim of the policy is to enable a truly connected, language-crossing Digital Single Market.
Economic growth and job creation	High	Thanks to this policy, many European SMEs will be enabled to participate in the multilingual DSM, substantially multiplying their reach and business. We expect the creation of tens of thousands of sustainable new jobs in the medium to long-term.
Tackling inequalities	High	Such a policy will facilitate efficient technology transfer from languages with richer resources and tools, to less-resourced languages.

4.4.2 Promote the European LT Platform of data and services

It is necessary to draw upon existing infrastructures and platforms to promote an open cloud-based platform, enabling the sharing and further development of LT-related resources developed at the European, national, and regional levels. The platform should also provide highly scalable, high-performance and robust basic tools for several LT applications.

Existing resource-exchange infrastructures such as META-SHARE and CLARIN have shown the necessity of storing and sharing linguistic data and resources in a single place. These existing infrastructures can play an important role in the construction of a European LT platform.

It is worth mentioning the platforms already developed in the Human Brain Project (HBP). The HBP has made extensive research in the deep learning and neural network areas, which are also the main current trends in HLT. There are two subprojects within the HBP that could benefit HLT, particularly the SP7 (High Performance Analytics and Computing Platform) and the SP9 (Neuromorphic

Computing Platform). Both provide platforms that could be used by LT researchers and industry leaders.

Because of the diversity of the LT applications and services, it may be more flexible to actually talk of connected existing platforms instead of a single platform, at least in a first phase, but it is important that they are conceptualized as a single European-based infrastructure for data and services. Moreover, a review of the key technology components required in the platforms is mandatory. Those key technology components for all EU languages that still have not been developed should be built and further integrated into the platform.

Also, regulation of the use of such data should be made much more open, and core language resources (annotated corpora, lexicons and ontologies) should be standardised and shared in this open environment. The regulation about crawling data should be carefully updated. Currently this area is a legal mine-field, but making use of data sets collected from the web for the purpose of building language technologies is crucial for the development of these technologies.

Evidence and assessment

Most improvements in HLT rely particularly on the ability to access and maintain ever larger and more finely tuned linguistic data. Lack of access to that data constrains development of language technologies. Innovation in HLT crucially depends on language resources, but currently there are not enough available resources to satisfy the needs of all languages, quantitatively and qualitatively. Language resources should be produced and made available for every language, every register, and every domain to guarantee full coverage and high quality tools. Acquiring and using data may rely on cooperation between the LT industry and the different constituencies that own, need and use these data. Collaboration between the industry and data owners will be needed. To put an example forward, public translation memories play an important role in the improvement of machine translation because they can be used to generate high quality training corpora. This also reduces development costs for companies because they can reuse an ever growing baseline corpora that many parties feed into and can focus on collecting the high value information that is specific to their client's project. The ability to share translation memories is also important for poorly resourced languages, for which there are often smaller batches of translations stored in many different locations, typically language service providers. If these can be pooled and shared, this will make it much easier for companies to create high quality translation engines for secondary language pairs.

The European academic and industrial language technology community is fully aware of the need for sharing resources such as language data, tools and core technology components as a basis for the successful development, implementation and continuous improvement of the multilingual services and applications. Initiatives such as FLaReNet and CLARIN have prepared the ground for a culture of sharing. Services such as META-NET's open resource exchange infrastructure, META-SHARE, can provide the basic technological platform as well as legal and organisational schemes. This effort shall revolve around the following axes: infrastructure; coverage, quality, and adequacy; language resources acquisition; openness; and interoperability. Interoperability of methods and services need to be guaranteed by significantly boosting standardisation activities already during the design phase. The goal is to base the platform and repositories 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.

Through the LT platform, researchers will be able to test and benchmark their technologies, providers will be able to offer their integrated services, and consumers will access services without having to install any software. It is worth noting that through the involvement of users, valuable data can be collected within these inherently European platforms.

Because of the demanding requirements regarding performance, reliability, user support, scalability, and persistence together with data protection and compliance with privacy regulation, it is desirable

that consortia with strong commercial partners participate in the establishment and operation of the platform(s).

Table 9: Assessment matrix for the "Promote the European LT Platform of data and services" policy

Criteria	Adequacy	Argument
Costs and benefits	High	The accumulation and sharing of linguistic data and resources in a single place may lower the R&D costs for new applications, in new languages and domains.
Political feasibility	Medium	The EU has started several initiatives in the line of unifying access to resources, e.g. EU Open Data Portal, the Connecting Europe Facility. Moreover there are several platforms that should be the basis of the HLP platform such as META-NET, CLARIN and the platforms developed by the HBP. Such an integrated LT-platform will build on these initiatives. However the scope of the action involving private as well as public actors may be hampered by a number of obstacles.
Feasibility in the European context	High	EU-funded projects such as FLReNet, CLARIN, HBP and notably META-SHARE have prepared the ground for a culture of sharing.
Effectiveness	High	A common European LT platform has the potential to facilitate the relationship between LT providers and LT consumers.
Sustainability	Medium	Sustainability covers preservation, accessibility, and operability (among other things). Collecting and preserving knowledge in the form of existing resources should be a key priority. A sustainability analysis must be part of a resource specification phase. Funding agencies should make a sustainability plan mandatory for projects concerned with the production of language resources.
Risks and uncertainties	Medium	The construction of the platform will start by a concerted harmonisation of several platforms, following well-grounded standardisation strategies. As pointed out, a concerted action of public and private actors will not be without its challenges.
Coherence with EU objectives	High	The objectives of the policy are aligned with EU objectives.
Potential ethical, social, environmental and regulatory impacts	High	There is a strong trend towards open data, i. e., data that are easily obtainable and that can be used with few, if any, restrictions. Sharing data and tools has become a viable solution towards encouraging open data.
Innovation	High	Europe needs a shared language infrastructure to underpin innovation in the LT industry.
Single Market	High	LTs services and data are key for the success of the Digital Single Market.
Economic growth and job creation	High	The platform will create more business to HLT providers and will facilitate access to specialised solutions to the European consumer, thereby accelerating economic growth and job creation.

Tackling inequalities	High	Europe needs a shared language infrastructure to preserve its languages, and give people and organisations access to the Digital Single Market using their own languages.
-----------------------	------	---

4.4.3 Bridge the technology gap between European languages

In order to bridge the technology gap, policies should focus on fostering basic research, applied research and development and innovation for language technologies for all European languages. Research and technology transfer among languages, along with increased collaboration across linguistic communities must receive more attention. Funding schemes, such as Horizon 2020 and the successor funding programme, could boost knowledge and technology transfer between countries and languages that already have excellent research and innovation hubs in LT and those that do not; the goal would be to enable the less resourced languages, including sign languages, to benefit from technologies already developed for other languages.

Evidence and assessment

Europe is a multilingual society. All official languages of the European Union enjoy, at least according to the founding documents of the EU, equal status. Nevertheless, more than 20 European languages are in danger of digital language extinction. Language technologies may serve as a vehicle for the protection and promotion of smaller and non-official languages. At present, LT is primarily used in relation to national and large regional languages, partly due to the investment required. However, to preserve the historical cultural and linguistic diversity of Europe and to facilitate an active participation of all European citizens in our democratic processes, it is also important for the smaller languages in Europe to make use of language technologies.

Relevant surveys, such as the META-NET White Paper Series “Europe’s Languages in the Digital Age”, reveal striking differences in LT support among the different European languages. While there are good-quality software and resources available for a few languages and certain application areas, other, usually smaller, languages show substantial gaps. Digital support for 21 of the 30 languages investigated in the aforementioned survey was “non-existent” or “weak” at best.

Certainly, the amount and quality of technologies available to a certain language tends to correlate with the number of speakers of that language. Among other reasons, companies refrain from investing in the development of sophisticated language technologies for languages spoken by a small number of speakers, thereby deepening the gap between well-supported languages and under-resourced ones. Moreover, not all countries have the required expertise or human resources necessary to support their languages. European policies must address this deficit.

Although much is said by the European institutions about the importance of linguistic diversity, very few policy initiatives are undertaken and even less funding is provided to support European linguistic diversity. They should aim to highlight this deficiency and promote the need for more support for all indigenous languages of Europe to ensure that our rich landscape of languages, many of them highly endangered, survive well into the future.

Table 10: Assessment matrix for the "Bridge the technology gap between European languages" policy

Criteria	Adequacy	Argument
Costs and benefits	Medium	An increased LT-support to minority languages has costs that may be lessened by promoting technology transfer across languages and countries, but it has the clear benefit of making the digital experience more equitable to the European citizens, irrespective of their language.
Political feasibility	Medium	Funding schemes, such as Horizon 2020, should be used for the purpose of technology transfer and bridging the gap across European languages.
Feasibility in the European context	High	Multilingualism is one of the cultural cornerstones of Europe. Contributing to language equality should be a priority of European public administrations.
Effectiveness	Medium	In addition to the EC / EU initiatives, it needs to be fully supported at the national level (Member States, Associated Countries, regions) for the spirit of this policy to succeed.
Sustainability	Low	Language technology transfer can help bridge the current gap among English and bigger languages on one side and minority languages on the other, but there will always need to be a sustained effort to keep the gap at bay.
Risks and uncertainties	High	Due to the many actors involved at the international, national and regional levels, risks and uncertainties are unavoidable, for certain languages more than for others.
Coherence with EU objectives	High	Fair access of all European citizens to the DSM irrespective of their native languages is an EU objective. LTs in all languages, no matter how small, is the key to achieve this objective.
Potential ethical, social, environmental and regulatory impacts	High	European citizens speaking a smaller language will have more opportunities to use their own language, e.g. when completing bureaucratic formalities or conducting business on line.
Innovation	Uncertain	Technology transfer in principle does not entail innovation in the research sense of the word, but it certainly may involve innovation in the local markets dimension.
Single Market	High	A true DSM in Europe cannot be based on a lingua franca or a reduced set of big languages, but needs to embrace each and every one of its languages.
Economic growth and job creation	High	Increased LT-support for all languages in Europe will benefit e-commerce activities, leading to economic growth and job creation.
Tackling inequalities	High	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, some of them actually facing the danger of digital extinction. Such a policy will reduce that threat.

4.5 Industry policies

4.5.1 Foster and support the development of investment instruments and accelerator programs targeting HLT start-ups

One of the biggest issues that companies must address when bringing their innovations to the market is the lack of private financing. While the research process is usually supported by public funding, both at the national and European level, the following steps in the business life cycle (development of the product, commercial activities, etc.) lack a coordinated public-private support in terms of funding. This fact is particularly relevant in the HLT sector. Several factors including the complexity of the technology, low understanding of the value of LT at the integrators level and the potential limited market when products and services are focused on a smaller language can deter private investors from spending on the HLT sector.

Many investment instruments and accelerator programs, which are supported and funded by European authorities, are focused on concrete technological topics and sectors. This approach allows adapting the characteristics of the programs to better fit the particularities of each technology/sector. The following list shows some examples:

- ODINE and FINODEX: open data.
- INCENSE and Climate-KIC Accelerator: clean tech.
- SOUL-FI: smart urban life (mobility, tourism, quality of life).
- FI-ADOPT: learning/training, healthy behaviour, social and cultural integration.
- EuropeanPioneers: media and content sector.
- FI-C3: smart territories, media & contents, care & well-being.
- EIT Health Accelerator: health tech.

Currently there are no specific accelerator programs targeting HLT start-ups. The strategic relevance of HLT for reinforcing the European common identity and citizenship engagement, as well as the preservation of the cultural heritage of all languages, including the smaller ones, and the effective development of the Digital Single Market requires a specific approach of investment instruments for all stages (seed capital, early stage, expansion stage, etc.).

Evidence and assessment

The European HLT industry is mainly made up by innovative smaller companies and micro-enterprises. Although most of them have been established in the market for several years, the specificities of the LT market in Europe (local/national companies with expertise in local languages that serve local markets) hamper their growth. The transformation into global players capable of competing with non-European big companies requires financing in all stages of business life cycle, not only in research activities. The creation of investment instruments and accelerator programs can increase the economic potential of the high level of innovation within the European industry, which could remain European instead of becoming subsidiaries of the USA's big players.

Table 11: Assessment matrix for the "Foster and support the development of investment instruments and accelerator programs targeting HLT start-ups" policy

Criteria	Adequacy	Argument
Costs and benefits	Medium	The creation of European public or mixed investment instruments might be costly but would have a positive direct impact in the LT Industry.
Political feasibility	High	Politicians of all EU countries would be willing to support this initiative, as it would allow to consolidate a key strategic industry for language diversity in Europe.
Feasibility in the European context	Medium	The investment instruments should be carefully defined in a way that do not entail State aid under Article 107(1) of the Treaty.
Effectiveness	High	The impact of increasing investment would be quantifiable in terms of employment created, growth of incomes and exports.
Sustainability	Medium	Companies will require more investments as they grow and the instruments defined may not cope with the growing demand of financing. The companies should seek other financing mechanisms, such as series B and C funding (from millions to hundreds of millions), IPO, etc.
Risks and uncertainties	Low	The volatility of the market and the threat of a strong competition from US players increase the risk of failed investments.
Coherence with EU objectives	High	The support of the European LT industry by means of financing can contribute to the creation of a truly Digital Single Market for all citizens and companies.
Potential ethical, social, environmental and regulatory impacts	Low	There are already several investment instruments complying ethical, social, environmental and regulatory requirements which may be adapted for the LT industry.
Innovation	High	The creation of investment instruments can benefit the innovation and the technology transfer from research institutions to the market.
Single Market	High	Increasing the investment for European LT companies may allow the rapid development of new products and services focused on facilitating business between EU countries and overcoming language barriers.
Economic growth and job creation	High	The creation of investment instruments may have a direct and positive impact on economic growth and job creation in the European LT industry.
Tackling inequalities	High	Language inequalities in the digital sphere can be conveniently addressed, strengthening the European LT industry by means of financing. While big US players may not be interested in supporting smaller languages, as it may not be profitable investing in them, European companies, strengthened by financial support, would be in a position to develop products and services for all European languages.

4.5.2 Increase the availability of qualified personnel on HLT

One of the biggest issues that the digital economy is facing in Europe is the shortage of technical professionals, which may endanger the economic growth expected through the intensive use of digital technologies in all productive sectors. This fact is more evident in the HLT European sector, as it is at the forefront of research and innovation and must face the fierce competition of US players, not only in the market but also in the recruitment of qualified professionals.

The ones with high potential, the very intelligent graduate students and PhDs with degrees in Computer Sciences, Computational Linguistics, AI, and Cognitive Sciences need incentives to stay in Europe.

European policies on this topic should be oriented towards:

- incentive models to retain talent in Europe. Establishing a start-up company should be made very easy and students need to be educated and informed about what it means to be an entrepreneur. There should be incentives to make successful and promising new HLT companies stay in Europe;
- analysing the current needs of LT-related education, together with research and industry, in order to design training programmes (formal and non-formal) for existing professionals and adapted or new courses of study in vocational and university education for students;
- raising awareness among students about the career opportunities in the HLT industry, with the aim of encouraging them to focus their studies on technical disciplines related to LT.

Evidence and assessment

Two indispensable prerequisites for innovative research and technology development are highly qualified researchers and software developers. During the preparation of the META-NET Strategic Research Agenda (SRA), many LT companies were approached. With almost no exceptions, the industry representatives mentioned the lack of qualified personnel to be a significant problem for their further growth and diminishing factor for producing innovative technologies. The SRA advises that Europe's academic programmes in natural language processing, speech processing, computational linguistics, language technologies, etc., should be further strengthened and advertised at an international level and made more attractive for potential students. The lack of skilled personnel currently is a bottleneck for many small and medium companies as well as for research centres.

Table 12: Assessment matrix for the "Increase the availability of qualified personnel on HLT" policy

Criteria	Adequacy	Argument
Costs and benefits	Medium	Although the cost of the proposed policies can be high, particularly raising awareness among students, the potential benefits for the industry having a wide pool of qualified professionals may compensate the public effort.
Political feasibility	Low	Member States are responsible for education policies. It can be difficult to include this issue in the educational priorities of all EU countries.
Feasibility in the European context	High	Promoting digital skills and professionals is a policy perfectly aligned with the development of the Digital Single Market.
Effectiveness	Medium	The effectiveness of this policy can only be measured in the medium/long-term, as it is not going to produce immediate impacts in the industry.

Sustainability	Low	The success of this policy lies in its maintenance over time. It would notably increase its cost.
Risks and uncertainties	Medium	The launch of this policy requires the participation of a great number of stakeholders: European institutions, Member States, academic institutions, industry, etc. The coordination of all stakeholders can be difficult and the objectives and priorities can be divergent among them.
Coherence with EU objectives	High	The proposed policy is coherent with the diverse initiatives launched by the European Commission in the field of digital skills and tech job promotions.
Potential ethical, social, environmental and regulatory impacts	Medium	The proposed policy would have a positive social impact, as it would contribute to the reinforcement of the European LT industry with qualified professionals, allowing greater attention to be directed towards regional and minority languages.
Innovation	High	The promotion of qualified professionals would improve the innovation in the sector, allowing ease in addressing new research challenges.
Single Market	High	The existence of more qualified professionals would allow LT companies to address the development of new products and services aimed at covering the needs of European SMEs related to cross-border business.
Economic growth and job creation	High	This policy would contribute to the economic growth of the LT industry, as well as to the creation of new, highly qualified jobs.
Tackling inequalities	High	The promotion of qualified professionals could improve the development of products and services for smaller languages, reducing the gap between them and English regarding their LT support.

4.6 Market policies

4.6.1 Raise awareness of the benefits for companies, public bodies, and citizens of the availability of on-line services, contents and products in multiple languages

The European society as a whole must be aware of the challenges of the multilingual Europe and how the fast improvement of HLT can effectively help to overcome language barriers. The main targets of this policy are: (1) the public bodies, namely policy-makers and public officials; (2) the firms, particularly SMEs that could benefit from access to a pan-European market thanks to an integrated DSM without language borders; and (3) the citizens that can benefit from multilingual public services and private offerings in a Europe without language barriers regardless of the languages they speak. It is crucial to show European stakeholders the benefits that these technologies bring to their businesses and daily lives, namely: accessing a bigger market for suppliers and consumers, and having the possibility to access public services and contents of other countries in the mother tongue of the speaker.

Evidence and assessment

Demand side policies intend to increase the demand of products and services that are available in the market and that have high potential for positive externalities, such as HLT. HLT are evolving and

improving very quickly while public bodies and firms are not drawing upon the benefits of this technology.

We believe that the low penetration of multilingual services in the public sector and the low availability of retailer's web pages³⁹ in several languages is related, among other causes, to a lack of knowledge and awareness of the rapid advancements of these technologies, their availability and how accurate they are becoming. Moreover, when analysing cross-border e-commerce we have found that it is very likely that people are not aware of the highly competitive offerings of products that are available in other countries' shops because they never reach those pages due to the "invisible" language barriers (they aren't aware of what they do not know). Therefore raising awareness about the availability and benefits of these technologies can substantially increase the demand and, along with public side policies, speed up the development and adoption of these technologies in the EU.

Table 13: Assessment matrix for the "Raise awareness of the benefits for companies, public bodies, and citizens of the availability of on-line services, contents and products in multiple languages" policy

Criteria	Adequacy	Argument
Costs and benefits	Medium	Raising awareness policies may be expensive and the benefits difficult to quantify.
Political feasibility	Medium	These policies should be defined carefully to send the right messages and not generate political opposition.
Feasibility in the European context	Medium	Designing a policy targeting the 28 EU countries is a challenging task. It is important to rely on national and regional governments that can be particularly interested in overcoming language barriers.
Effectiveness	Medium	The effect of these policies is difficult to quantify.
Sustainability	Medium	If the policy does not have short term results it can be difficult to keep it sustained over time.
Risks and uncertainties	Medium	There is no magic recipe for raising awareness and therefore the result of the policy is subject to medium levels of uncertainty.
Coherence with EU objectives	High	Creating a fair and integrated Europe is one of the main targets of the EU. Therefore, showing society how technology can help integrate Europe while preserving its cultural diversity is fully aligned with EU objectives.
Potential ethical, social, environmental and regulatory impacts	Medium	Sending the right messages should not have any negative implication.

³⁹ In 2012, 82 % of European online shopping sites were published in a single language; 11 % were published in two languages and only 2 % offered five or more languages, without any site that could be considered fully multilingual (LT-Innovate, 2012a).

Innovation	High	By increasing the demand it is expected that a healthier and more innovative market arise.
Single Market	High	Showing society how one of the main barriers for creating a DSM is overcome by using innovative technologies will contribute to sharing the benefits of the Single Market among citizens and firms.
Economic growth and job creation	Medium	If the policy is successful it can contribute to direct and indirect economic growth by promoting a more competitive market.
Tackling inequalities	High	Those countries and groups of citizens that are more likely to be left behind by language barriers are the main target of the policy and therefore will benefit the most.

4.6.2 Promote the automated translation of e-commerce websites of European SMEs

Within the SMEs, wholesale and retail are two of the most crucial sectors that could particularly benefit from increasing their market size by providing their services in multiple European languages. However, language barriers and technical issues are hindering European SMEs web merchants from selling to other countries. Therefore we propose providing economic incentives to: (1) SMEs, to translate their web-based e-shops using HLT; and (2) to ICT integrators, to develop specific cloud based services allowing a smooth integration of HLT in the e-commerce applications used by SMEs.

There is a substantial amount of structural funds to increase the competitiveness of the SMEs that can be used to fund this policy. In fact, countries receiving a higher amount of European Structural and Investment Funds for the period 2014-2020 for the theme “Competitiveness of SMEs” have a higher percentage population of people who are unable to speak any of the three European major languages as described in Annex 7.8. SMEs in those countries are likely to benefit most from using HLT.

Evidence and assessment

SMEs are a crucial pillar of the European economy⁴⁰ and the effect of not having adequate language skills within the SMEs can be particularly challenging for Europe. Wholesale and retail trade is the most important sector for microenterprises and SMEs in terms of employment, value added and number of enterprises (Muller et al., 2014). This sector, therefore, is one that could benefit more from cross-border transactions. For SMEs, however, selling on-line abroad is more challenging than for larger companies when considering language barriers. In 2015, 23 % of large companies sold through electronic channels to other EU countries compared to only 12 % and 7 % of medium and small companies, respectively. While larger web-merchants are likely to have the resources to sell internationally, this is not the case for smaller companies.

Although we acknowledge that there are many factors hindering SMEs from selling to other countries – market uncertainties, legal issues, logistic barriers, geo-blocking – we still believe that one of the main reasons for this gap is the language barrier. Using the Community Survey on ICT Usage and e-Commerce of 2009, Pavel (2010) found that the only significant difference between barriers hindering

⁴⁰ SMEs represent around 99.8 % of all enterprises and account for around two-thirds of total employment in 2014 (71.4 % of the increase in employment in that year) in the non-financial business sector (Muller et al., 2014).

electronic sales from large enterprises on one side, and small and medium enterprises on the other, were the language problems and technical issues. Regarding the language barriers, it is expected that efficient HLT solutions will help foster an increasingly level playing field, where small European retailers could compete more easily with big web-stores, many of them in non-European hands.

Regarding the technical issues, one of the last ICT innovations that can help SMEs to increase their productivity and competitiveness is the cloud computing paradigm. However, the lack of cloud solutions in local languages might hinder the adoption of these technologies by SMEs in the short-term (Bradshaw, D. et al., 2012).

High quality HLT services adapted to the specificities of the SMEs could be the answer by providing feasible, quick and efficient solutions to these companies.

Table 14: Assessment matrix for the "Promote the automated translation of e-commerce web-sites of European SMEs" policy

Criteria	Adequacy	Argument
Costs and benefits	High	The policy can be funded by using current structural funds for the "Competitiveness of SMEs".
Political feasibility	High	Policies affecting SMEs are usually very welcome.
Feasibility in the European context	High	There are already policies and structural funds to increase the competitiveness of the SMEs, so it is a topic that is already in the European policy agenda.
Effectiveness	Low	The effect of policies to finance investment of SMEs is not easy to assess. It is expected that if SMEs see a clear benefit on a specific investment they would likely make it by themselves.
Sustainability	Medium	May be unfeasible to sustain a policy like this over time. However, if the policy yields positive results it is expected that other SMEs will follow without requiring public funding
Risks and uncertainties	Low	There is no magical recipe for incentivising SMEs.
Coherence with EU objectives	High	SMEs are a crucial pillar of European economy and therefore it is coherent that the EU support these companies.
Potential ethical, social, environmental and regulatory impacts	Medium	No concerns regarding this topic.
Innovation	Medium	No mayor innovations are expected apart from the integration of innovative technologies in the SMEs.
Single Market	High	The impact on the cross-border e-commerce in the DSM is expected to be high if the rate of European SMEs selling to other countries substantially increases.

Economic growth and job creation	Medium	If the policy is successful it can contribute to direct and indirect economic growth by promoting more competitive SMEs.
Tackling inequalities	High	It could benefit particularly SMEs in countries with small languages.

4.7 Public service policies

4.7.1 Public procurement of innovative technology and pre-commercial public procurement

Public procurement of innovation (PPI) has been recognized by the European Union as an important tool to foster innovation, reduce the time new solutions arrive on the market, improve the quality of public services and increase the competitiveness of European enterprises, particularly SMEs. Its relevance has been highlighted by the Europe 2020 flagship initiative Innovation Union. Public administrations need to ask for ground-breaking technologies, which are then developed by R&D&I in a longer time-frame.

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 (the public administration) and the supplier share risks and benefits under market conditions.

The application of PPI and PCP when implementing multilingualism across public services, contents and products in the Union such as intelligent assistants for migrants, translating services in Justice and Health, multilingual e-government services and legal documents available in all European languages, can become a strategic step in fostering the development and growth of the HLT industry in Europe.

Evidence and assessment

Implementing PPI and PCP for the acquisition of HLT by public administrations in Europe can improve the competitiveness of the sector and stimulate the offer of these products, particularly those from SMEs, as well as reduce the time to market of new products and services. The public administration, as a lead customer, increases the demand and encourages the creation of new businesses. This increasing competition is expected to result in a greater economy, efficiency and effectiveness, as well as fostering innovation. Finally, it will help face the societal challenge of multilingualism in a more efficient way.

Encouraging PPI and PCI of HLT can boost the quality and availability of these technologies that will result in better multilingual public services, both for the public and private sectors, and therefore reducing the current fragmentation of the Digital Single Market. Information and services could be then provided in any language of the European Union, which will reduce inequalities among languages and stimulate the mobility of businesses, citizens and workers in Europe.

Table 15: Assessment matrix for the "Public procurement of innovative technology and Pre-commercial public procurement " policy

Criteria	Adequacy	Argument
Costs and benefits	Medium	The implementation of PPI and PCP instruments for HLT might be more costly in the short run, but it would reduce translation costs in the medium term while expanding to all languages services and information that currently are only available only in one or two languages.
Political feasibility	High	PPI and PCP have already been included in the Europe 2020 strategy, and 2014 EU procurement directives allow for its application.
Feasibility in the European context	High	Boosting these innovative ways of procurement is completely aligned with the development of the Digital Single Market and the Europe 2020 objectives.
Effectiveness	Medium	The effectiveness of this policy can only be measured in the medium/long-term, as it is not going to produce immediate impacts in the industry.
Sustainability	Medium	The quick evolution of technologies might require a continuous update of products and services.
Risks and uncertainties	Medium	PPI and PCP, particularly the former, entail a certain amount of risk. Pilot projects can help reduce those risks.
Coherence with EU objectives	High	The proposed policy is coherent with the Europe 2020 strategy.
Potential ethical, social, environmental and regulatory impacts	Medium	The proposed policy would have a positive social impact, as it would contribute to the reinforcement of the European LT industry and will allow for real multilingualism of public services around Europe.
Innovation	High	PPI and PCP are tools for boosting innovation.
Single Market	High	Higher competitiveness of the European industry would allow LT companies to address the development of new products and services aimed at covering the needs of European SMEs related to cross-border business. Additionally, the expansion of multilingual public services will facilitate the mobility of business, workers and citizens.
Economic growth and job creation	High	This policy would contribute to the economic growth of the LT industry, as well as to the creation of new qualified jobs
Tackling inequalities	High	By acquiring innovative HLT public authorities will be able to provide multilingual public services and information.

4.7.2 Foster the translation of national and regional public web-sites and documents to other EU languages by using HLT

To create a truly integrated Europe it is important that EU citizens and businesses can access relevant public information and use the public online services of any of the Member States in their own language. Public digital services throughout Europe cannot be truly effective, efficient, citizen-centric and inclusive in the current context of language fragmentation. Public administrations at all levels: local, regional, national and European, should be encouraged to use HLT and provide information and services in all European languages.

Additionally, the availability of public services around Europe in all languages would contribute to increasing citizens' and civil society's scrutiny and engagement, and to further improving trust in public bodies.

Having legal support to promote translation of public websites to all European languages could speed up the deployment of this policy. This policy could be enforced by establishing a legal mandate or recommendation to national and regional governments to provide their web pages in several languages when receiving structural funds. As a first goal, we encourage that certain relevant public services in all countries must be available in, at least, the official EU languages.

The action "Overcoming language barriers" of the ISA² program can be a good starting point, although it should be substantially reinforced and managed by a specific unit within EC dedicated to multilingualism.

Evidence and assessment

There are three main types of target groups that are especially interested in multilingual public services: a) international travellers, business people and tourists, b) migrant groups and c) speakers of regional and smaller languages. However, public bodies usually provide information in very few (if any) languages different to the official language of the country. For instance, in considering a sample of European cities, English was found to be the main non-official language next to the national language, followed by German and French. Multilingual public services are usually related to tourism, immigration and integration, legal services and transport while cultural and citizenship participation rank lower in multilingual services (Extra, Guus & Kutlay, Yağmur, 2012). In the analysis, we have seen that 34 % of public administrations' portals of Member States only provided information in the official languages of the country, and 66 % offered information in other languages, with English being the most common (61 %).

Currently, the project MT@EC based on the open source software Moses allows machine translation between the 24 EU official languages, and can be used by European institutions and public administrations in EU countries. The Connecting Europe Facility initiative is expected to provide a Service of Automated Translation (CEF.AT); however, the current budget of CEF.AT is estimated at 4 million euro (2014) and 8 million euro (2015) (European Commission, 2015d) and might not be enough to achieve its targets.

The policy may be complemented by providing incentives to national and regional governments through structural funds. Countries receiving a higher amount of European Structural and Investment Funds for the period 2014-2020 for the theme "Efficient Public Administration" have a higher percentage population of people who are unable to speak any of the three European majority languages, as described in Annex 7.8. Those countries are likely to benefit most from using HLT.

Table 16: Assessment matrix for the "Foster the translation of national and regional public webs and documents to other EU languages by using HLT" policy

Criteria	Adequacy	Argument
Costs and benefits	Low	From the purely economic point of view, the cost of offering public information and services in all European languages through HLT will be high and the benefits will be mainly intangible and include more equality, more integration, etc.
Political feasibility	Medium	There may be reluctance on the part of national or local politicians or civil servants.
Feasibility in the European context	High	Only at the European level these policies could be implemented. Existing instruments such as Structural Funds could be used for this policy.
Effectiveness	Low	The real usage of local public services of a Member State in languages from other Member States will probably be limited. However, it might have an important impact on awareness and symbolic effect.
Sustainability	Medium	The success of this policy lies in its maintenance over time, what would notably increase its cost.
Risks and uncertainties	Low	Risks would only be related to the maturity of the technology.
Coherence with EU objectives	High	The construction of a European identity should rely on the protection and promotion of European languages, as well as the inclusion of all citizens in a real Single Market.
Potential ethical, social, environmental and regulatory impacts	Low	No ethical, social, environmental or regulatory impacts are expected.
Innovation	High	The use of HLT by national and regional governments can help boosting innovation in HLT at the local levels and encouraging the industry.
Single Market	High	Cross-border, multilingual, citizen-centric services can help achieving a Single Market.
Economic growth and job creation	High	This policy will stimulate the market and increase competition, which will have a positive effect on the industry and the sector.
Tackling inequalities	High	All citizens will be able to access any public information or service in the EU, thereby reducing inequalities. Additionally, this will allow comparisons of services and increase the accountability of public authorities.

5 Conclusions

The EU is a unique endeavour involving more than 500 million citizens sharing about 80 different languages. While multilingualism is one of the biggest assets of Europe, it is also one of the most substantial challenges for the creation of a truly integrated EU. On the one hand, European diversity is very valuable for the European society, as shown by the huge amount of contents and services available in the multiple languages of our continent. On the other hand, as shown in Table 17, language barriers among countries, citizens, and businesses have strong social and economic consequences such as (1) fostering a language divide, (2) hampering workers' mobility, (3) hindering access to cross-border public-services, (4) reducing citizens' engagement and participation in political debates and processes, and (5) creating fragmented markets of cross-border trade and e-commerce. Overcoming language barriers represents both a crucial opportunity and a formidable challenge for the EU.

Table 17: Summary of the main socio-economic consequences of language barriers analysed in the study

Impact	Unit	Estimated figure	Source
Population unable to speak English	Percentage of EU population	60.63 %	Compiled by the authors based on the Eurobarometer 77.1 (European Commission, 2014a)
Population unable to speak one of the six most spoken EU languages	Percentage of EU population	14.93 %	
Education gap of citizens speaking English as a foreign language	Increase in the chances of speaking English (leaving education after 19 compared to leaving education before 16)	19 more chances	
Age gap of citizens speaking English as a foreign language	Increase in the chances of speaking English (younger than 30 compared to older population)	5 more chances	
Effect of language barriers on European mobility	Increase in the number of people moving between EU countries with low language barriers to live and work compared to countries with high language barriers	3.25 times higher	Compiled by the authors based on (European Commission, 2009, 2014a; EUROSTAT, 2016c)
Effect of language barriers on cross-border e-commerce	Increase in the number of cross-border e-shoppers between countries with low language barriers compared to countries with high language barriers	4.14 times higher	Compiled by the authors based on (Civic Consulting, 2011; European Commission, 2014a; EUROSTAT, 2016c)

No feasible solution has been provided so far. In considering multilingualism policies there is always a trade-off between effectiveness and fairness; between utopia and reality. Europe has several choices: either (1) we all learn one lingua franca (for example, English) – this is not only very unlikely, but it would be very much against the European founding idea, by which all its languages enjoy the same

status; alternatively (2) we learn additional languages – which is better, but also more unfeasible and still would not solve the problem because one cannot learn all required languages; or finally (3) we draw upon human language technologies to respect the right of all European citizens to be able to communicate using their own languages. The truth is that the EC itself has opted for the first choice by moving towards a monolingual scenario where, in practice, English has become the only working language.

Human language technologies become a crucial technology for the construction of a fair and integrated EU. The growth and improvement of data-based, machine-learning translation technologies are making these technologies a real solution to overcoming language barriers. The increasing availability of high quality translation and real-time speech recognition services are creating the tools for businesses and public bodies to seize the opportunities to provide their services, contents and support in any language.

However, the European LT research and industry community is not able to effectively respond to this challenge, while top players, such as Google or Microsoft, are not European and therefore may be unfit to address the specific needs of Europe. Although smaller languages are the ones to gain most from these technologies, tools and resources for them are scarce, in some cases almost non-existent, and as a consequence, many European languages are in danger of digital extinction.

Moreover, language technologies currently do not play a role in the European political agenda and not properly reflected in current EU Information and Communication Technologies policies – all this despite the fact that, according to the founding documents of the European Union, all languages enjoy the same status.

Acknowledging that technology is the only possible way to address the multilingual challenge, we need to move swiftly, establishing the European Human Language Project as an instrument to boost Europe towards the next generation of HLT development and to address the issue of Europe's multilingualism. The Human Language Project consists of a set of policies involving multiple stakeholders at the European, national and regional levels in a coordinated and joint effort to move Europe into the main position in this field so that it can develop and benefit from the next generation of language technologies.

We suggest that the EC should be in charge of the overall coordination of the policies of the HLP at the different levels. Not only European institutions, but also national and regional governments should be involved in creating resources for their own languages. Research in Europe should focus on creating the new paradigm of HLT, combining a fresh look at linguistics with the power of current artificial intelligence methods, which need vast knowledge bases, being themselves created with the help of sophisticated language technologies. Talent scarcity and brain drain should be transformed into talent creation and brain gain. Resources for the industry should be provided in a seamless, open and effective way through existing European interoperable platforms. The public sector should provide their contents and services for all European languages while promoting the growth of the HLT market through public procurement of innovative technology. Mechanisms to facilitate the scaling-up of European innovative HLT companies should be promoted. Eventually, policies supporting firms across Europe that sell cross-border by providing their contents, products, and services in the different European languages should be enacted to create a fully integrated DSM.

Within the Human Language Project 11 policies are proposed. These policies are structured into five groups: institutional policies, research policies, industry policies, market policies and public services policies.

Institutional policies involve initiatives to adapt current institutional frameworks to draw upon emerging technology trends to better fit the challenges of a multilingual Europe while properly assessing the results. Research policies focus on moving Europe towards the development of the next generation of Language Technologies. Research policies also aim at integrating research and industry, providing Europe with the tools to share resources to effectively compete with other markets. At the

same time, these tools will help contribute to the equality of all European citizens in their everyday digital experience regardless of their language. Industry policies foster the creation and growth of competitive European firms while increasing the availability of highly qualified workers. Market policies seek to improve the HLT sector in Europe by raising awareness among European stakeholders of the relevance of these technologies to further increase the demand of their services. There is a specific policy targeting small web merchants so they can benefit from accessing a much bigger market by translating their web shops using HLT. Public service policies intend to create multilingual public services in the European, National, Regional and Local administrations while contributing to increase the innovative HLT sector by using public procurement tools.

The policy options have been assessed using a multi-criteria weighted matrix. A summary of the policies, including the criteria assessment and the public and private stakeholders involved in the policies, is shown in Table 18. The policies are ordered by the overall assessment. We followed an approach where the different policy options reinforce each other instead of being mutually exclusive, so the overall assessment should be used to prioritize the policies and not to exclude some of them.

Table 18: Summary of the policy options of the HLP

	Proposed policies for the European Human Language Project	Criteria												Overall assessment (1 to 100)	Stakeholders involved					
		Costs and benefits	Political feasibility	Feasibility in the European context	Effectiveness	Sustainability	Risks and uncertainties	Coherence with EU objectives	Potential ethical and other impacts	Innovation	Single Market	Economic growth and job creation	Tackling inequalities		EP	EC	National Governments	Regional Governments	Research institutions	HLT Industry
		9	7	6	8	6	4	6	4	8	10	8	10							
	Criteria weight (1 to 10)																			
Research	Refocus and strength research in HLT	↑	↑	↑	↑	↑	↓	↑	↓	↑	↑	↑	↑	93						
Research	Promote the European LT Platform of data and services	↑	↓	↑	↑	↓	↓	↑	↑	↑	↑	↑	↑	90						
Institutional	Reinforce the role of HLT within the institutional framework of multilingualism related bodies	↑	↑	↓	↑	↑	↓	↑	↓	↓	↑	↓	↑	83						
Public services	Public procurement of innovative technology and Pre-commercial public procurement	↓	↑	↑	↓	↓	↓	↑	↓	↑	↑	↑	↑	82						
Industry	Foster and support the development of investment instruments and accelerator programs targeting HLT start-ups	↓	↑	↓	↑	↓	↓	↑	↓	↑	↑	↑	↑	78						
Research	Bridge the technology gap between European languages	↓	↓	↑	↓	↓	↑	↑	↑	↑	↑	↑	↑	76						
Market	Promote the automated translation of e-commerce web sites of European SMEs	↑	↑	↑	↓	↓	↓	↑	↓	↓	↑	↓	↑	71						
Industry	Increase the availability of qualified personnel on HLT	↓	↓	↑	↓	↓	↓	↑	↓	↑	↑	↑	↑	70						
Market	Raise awareness of the benefits for companies, public bodies, and citizens of the availability of on-line services, contents and products in multiple languages	↓	↓	↓	↓	↓	↓	↑	↓	↑	↑	↓	↑	70						
Public services	Foster the translation of national and regional public webs and documents to other EU languages by using HLT	↓	↓	↑	↓	↓	↓	↑	↓	↑	↑	↑	↑	63						
Institutional	Create tools to properly evaluate HLT policies	↓	↓	↓	↓	↓	↓	↑	↓	↓	↑	↓	↑	57						

6 References

- Achim Ruopp, & Jaap van der Meer. (2015). *Moses MT Market report*. Retrieved from http://www.statmt.org/mosescore/uploads/Internal/D4.8_Moses_MT_Market_Report.pdf
- Alexa. (2016). Alexa Top 500 Global Sites. Retrieved 22 September 2016, from <http://www.alexa.com/topsites>
- Auguie, B. (2016). *gridExtra: Miscellaneous Functions for 'Grid' Graphics*. Retrieved from <https://CRAN.R-project.org/package=gridExtra>
- Bivand, R., & Lewin-Koh, N. (2016). *maptools: Tools for Reading and Handling Spatial Objects*. Retrieved from <https://CRAN.R-project.org/package=maptools>
- Bojanowski, M. (2015). *intergraph: Coercion Routines for Network Data Objects*. Retrieved from <http://mbojan.github.io/intergraph>
- Bort, J. (2016, April 28). Google CEO's vision for the future sounds a lot like Microsoft's. Retrieved 14 November 2016, from <http://www.businessinsider.com/googles-vision-sounds-like-microsofts-2016-4>
- Bowen, S. (2015). *The impact of language barriers on patient safety and quality of care*. Société Santé en français. Retrieved from <http://santefrancais.ca/wp-content/uploads/SSF-Bowen-S.-Language-Barriers-Study.pdf>
- Bradshaw, D., Folco, G., Cattaneo, G., & Kolding, M. (2012). *Quantitative Estimates of the Demand for Cloud Computing in Europe and the Likely Barriers to Uptake*. IDC.
- British Council. (2011). Towards a language rich Europe: multilingual essays on language policies and practices. Retrieved from https://www.teachingenglish.org.uk/sites/teacheng/files/LRE_FINAL%20WEB.pdf
- Brownrigg, D. M. P. for R. by R., Minka, T. P., & Bivand, transition to P. 9 codebase by R. (2015). *mapproj: Map Projections*. Retrieved from <https://CRAN.R-project.org/package=mapproj>
- Cardona, M., Duch-Brown, N., Francois, J., Martens, B., & Yang, F. (2015). *The macro-economic impact of e-commerce in the EU Digital Single Market*. Institute of Prospective Technological Studies, Joint Research Centre.
- Civic Consulting. (2011). *Consumer market study on the functioning of e-commerce and Internet marketing and selling techniques in the retail of goods*. Retrieved from http://ec.europa.eu/consumers/archive/consumer_research/market_studies/docs/study_ecommerce_goods_en.pdf
- Civil Society Platform on Multilingualism. (2011, June 9). Policy Recommendations for the Promotion of Multilingualism in the European Union. Retrieved 17 October 2016, from http://ec.europa.eu/languages/information/documents/report-civil-society_en.pdf
- Council of the European Communities. (1991, July 29). Council Decision of 29 July 1991 on the introduction of a single European emergency call number. Retrieved from <http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:31991D0396&from=EN>
- Cracker, & LT_Observatory. (2015, April 22). Strategic Agenda for the Multilingual Digital Single Market. Retrieved from <http://www.meta-net.eu/projects/cracker/multimedia/mdsm-sria-draft.pdf>
- Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695.
- DG CONNECT. (2014). Inventory of the reports on the studies completed by the European Commission Directorate General for Communications Networks, Content and Technology (DG CONNECT) |

Digital Single Market. Retrieved 13 September 2016, from <https://ec.europa.eu/digital-single-market/en/news/inventory-studies-reports-procured-or-realised-european-commission-directorate-general>

DG Translation. (2009). *The size of the language industry in the EU*. Retrieved from http://bookshop.europa.eu/en/study-on-the-size-of-the-language-industry-in-the-eu-pbHC8009985/downloads/HC-80-09-985-EN-N/HC8009985ENN_002.pdf?FileName=HC8009985ENN_002.pdf&SKU=HC8009985ENN_PDF&CatalogueNumber=HC-80-09-985-EN-N

EC. (2015). Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: A Digital Single Market Strategy for Europe. Retrieved from http://ec.europa.eu/priorities/digital-single-market/docs/dsm-communication_en.pdf

Ecommerce Foundation. (2016). *Cross-border e-Commerce barometer 2016: Barriers to Growth - May 2016*. Ecommerce Europe. Retrieved from <http://www.ecommerce-europe.eu/stream/research-report-cross-border-e-commerce-barometer-2016.pdf>

ECSMP. (2016). Statutes. Retrieved from http://ecspm.org/wp-content/uploads/2016/04/ECSMP_-Statutes_draft.pdf

EFNIL. (2016). European Federation of National Institutions for Language. Retrieved 14 November 2016, from <http://www.efnil.org/>

Ethnologue. (2013). Ethnologue - Languages of the World. Retrieved from www.ethnologue.com

Eurobarometer. (2010). *The European Emergency Number 112. Analytical report. Wave 3*. European Commission. Retrieved from http://ec.europa.eu/public_opinion/flash/fl_285_en.pdf

Eurobarometer, S. (2012). 386 «Europeans and their Languages».

European Commission. (2015). *Studies on translation and multilingualism: Public service translation in cross-border healthcare*. DG Translation. Retrieved from http://ec.europa.eu/dgs/translation/publications/studies/summary_public_service_translation_healthcare_eu_en.pdf

European Commission. (2009). *Eurobarometer 72.5 (Nov-Dec 2009). TNS OPINION & SOCIAL, Brussels [Producer]. GESIS Data Archive, Cologne. ZA4999 Data file Version 5.1.0, doi:10.4232/1.11642*. Brussels. Retrieved from <https://dbk.gesis.org/dbksearch/sdesc2.asp?no=4999&search=mobility&search2=&DB=e&tab=0&no tabs=&nf=1&af=&ll=10>

European Commission. (2010). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions A Digital Agenda for Europe* (No. COM/2010/0245 final). Retrieved from <http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=COM:2010:0245:FIN>

European Commission. (2011). *Single Market Act Twelve levers to boost growth and strengthen confidence 'Working together to create new growth'* (No. 52011DC0206). Retrieved from <http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1477155902083&uri=CELEX:52011DC0206>

European Commission. (2012a). *Bringing e-commerce benefits to consumers* (COMMISSION STAFF WORKING DOCUMENT No. SEC(2011) 1640 final). Retrieved from http://ec.europa.eu/internal_market/e-commerce/docs/communication2012/SEC2011_1640_en.pdf

European Commission. (2012b). *COMMISSION COMMUNICATION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS A coherent framework for building trust in the Digital Single Market for e-*

commerce and online services (Commission Communication No. COM/2011/0942 final). Retrieved from <http://eur-lex.europa.eu/legal-content/hr/ALL/?uri=CELEX:52011DC0942>

European Commission. (2012c). *Flash Eurobarometer 331 (Retailers' Attitudes Towards Cross-border Trade and Consumer Protection, wave 2)*. TNS Political & Social [producer]. GESIS Data Archive, Cologne. ZA5614 Data file Version 1.0.0, doi:10.4232/1.11455. Brussels. Retrieved from <https://dbk.gesis.org/dbksearch/SDesc2.asp?ll=10¬abs=&af=&nf=&search=&search2=&db=E&no=5614>

European Commission. (2012d). *Flash Eurobarometer 332 (Consumers' Attitudes Towards Cross-border Trade and Consumer Protection, wave 2)*. TNS Opinion & Social [producer]. GESIS Data Archive, Cologne. ZA5615 Data file Version 2.0.0, doi:10.4232/1.11494. Brussels. Retrieved from <https://dbk.gesis.org/dbksearch/SDesc2.asp?ll=10¬abs=&af=&nf=&search=&search2=&db=E&no=5615>

European Commission. (2013a). *Council Decision (2013/743/EU) establishing the specific programme implementing Horizon 2020 | IPR Helpdesk* (Council Decision No. 2013/743/EU). Retrieved from <https://www.iprhelpdesk.eu/node/2206>

European Commission. (2013b). *Final Report of the WP2: Requirements and Scenarios of the DIGIT - Federated catalogue of public services*. Retrieved from https://ec.europa.eu/isa/documents/final-report-phase-2-requirements-and-scenarios_en.pdf

European Commission. (2013c). *Implementation of the European emergency number 112 - Results of the sixth data-gathering round*. Retrieved from http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=1674

European Commission. (2013d, December 10). *HORIZON 2020 WORK PROGRAMME 2014–2015; 5. Leadership in enabling and industrial technologies; i. Information and Communication Technologies*. Retrieved from http://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/main/h2020-wp1415-leit-ict_en.pdf

European Commission. (2014a). *Eurobarometer 77.1: Robotics, Civil Protection, Humanitarian Aid, Smoking Habits, and Multilingualism, February-March 2012 TNS OPINION & SOCIAL, Brussels [Producer]*. GESIS Data Archive, Cologne. ZA5597 Data file Version 3.0.0, doi:10.4232/1.12014. Brussels. Retrieved from <https://dbk.gesis.org/dbksearch/sdesc2.asp?no=5597&db=e&doi=10.4232/1.11481>

European Commission. (2014b). *Regulation (EU) No 283/2014 of the European Parliament and of the Council of 11 March 2014 on guidelines for trans-European networks in the area of telecommunications infrastructure and repealing Decision No 1336/97/EC Text with EEA relevance* (Regulation No. 32014R0283). Retrieved from <http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32014R0283>

European Commission. (2014c). *Translation and multilingualism*. DG Translation. Retrieved from <http://bookshop.europa.eu/uri?target=EUB:NOTICE:HC0414307:EN>

European Commission. (2015a). *COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS A Digital Single Market Strategy for Europe* (Commission Communication No. COM/2015/0192 final). Retrieved from <http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52015DC0192>

European Commission. (2015b). *Flash Eurobarometer 396 (Retailers' Attitudes Towards Cross-border Trade and Consumer Protection, wave 4)*. TNS Political & Social [producer]. GESIS Data Archive, Cologne. ZA5942 Data file Version 1.0.0, doi:10.4232/1.12118. Brussels. Retrieved from <https://dbk.gesis.org/dbksearch/sdesc2.asp?no=5942&search=online&search2=&DB=e&tab=0¬abs=&nf=1&af=&ll=10>

European Commission. (2015c). *Flash Eurobarometer 413 (Companies Engaged in Online Activities)*. TNS Political & Social [producer]. GESIS Data Archive, Cologne. ZA6284 Data file Version 1.0.0, doi:10.4232/1.12353. Brussels. Retrieved from <https://dbk.gesis.org/dbksearch/SDesc2.asp?ll=10¬abs=&af=&nf=&search=&search2=&db=E&no=6284>

European Commission. (2015d). *Machine Translation for Public Administrations (MT@EC)*. Retrieved from <http://ec.europa.eu/isa/documents/presentations/20151125-mtec-isa-workshop-pl.pdf>

European Commission. (2015e). The Digital Single Market Blog | Digital Single Market. Retrieved 28 September 2016, from <https://ec.europa.eu/digital-single-market/en/blog>

European Commission. (2015f). *The European Consumer Centres Network: 10 years serving Europe's consumers : anniversary report 2005-2015*. Luxembourg: Publications Office.

European Commission. (2015g, February 7). Language Technologies and Big Data | Digital Single Market. Retrieved 27 October 2016, from <https://ec.europa.eu/digital-single-market/en/language-technologies-and-big-data>

European Commission. (2015h, November 11). CORDIS - EU research projects under FP7 (2007-2013) - Datasets. Retrieved 11 November 2016, from <https://data.europa.eu/euodp/es/data/dataset/cordisfp7projects>

European Commission. (2016a). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. EU eGovernment Action Plan 2016-2020. Accelerating the digital transformation of government* (No. COM(2016) 179 final). Retrieved from <http://ec.europa.eu/transparency/regdoc/rep/1/2016/EN/1-2016-179-EN-F1-1.PDF>

European Commission. (2016b). DG-Translation organisation chart. Retrieved 18 October 2016, from http://ec.europa.eu/dgs/translation/whoweare/organisation_chart_en.pdf

European Commission. (2016c). EUROPA - Topics of the European Union – Multilingualism. Retrieved 17 October 2016, from https://europa.eu/european-union/topics/multilingualism_en

European Commission. (2016d). ISA2 work programme. 2016 financial overview. Retrieved from https://ec.europa.eu/isa2/sites/isa/files/library/documents/isa2-work-programme-2016-financial-overview_en.pdf

European Commission. (2016e). Language Technologies | Digital Single Market. Retrieved 19 October 2016, from <https://ec.europa.eu/digital-single-market/en/language-technologies>

European Commission. (2016f). Open Data Portal for the European Structural Investment Funds - European Commission. Retrieved 4 November 2016, from <https://cohesiondata.ec.europa.eu/funds>

European Commission. (2016g). Population data - Eurostat. Retrieved 27 October 2016, from <http://ec.europa.eu/eurostat/web/population-demography-migration-projections/population-data>

European Commission. (2016h). *Proposal of a regulation of the European Parliament and of the Council on addressing geo-blocking and other forms of discrimination based on customers' nationality, place of residence or place of establishment within the internal market and amending Regulation (EC) No 2006/2004 and Directive 2009/22/EC*. Brussels. Retrieved from http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=15952

European Commission. (2016i). Who we are | DG-Connect. Retrieved 18 October 2016, from <https://ec.europa.eu/digital-single-market/en/who-we-are-dg-connect>

European Commission. (2016j, May). Migration and migrant population statistics - Statistics Explained. Retrieved 27 October 2016, from http://ec.europa.eu/eurostat/statistics-explained/index.php/Migration_and_migrant_population_statistics

European Commission. (2016k, July). Boosting e-commerce in the EU | Digital Single Market. Retrieved 10 August 2016, from <https://ec.europa.eu/digital-single-market/en/boosting-e-commerce-eu>

European Commission. (2017). ISA² - Interoperability solutions for public administrations, businesses and citizens. Retrieved 3 February 2017, from https://ec.europa.eu/isa2/home_en

European Economic and Social Committee. (2011). *EESC Opinions. A Digital Agenda for Europe* (No. TEN/426 EESC-2010-1628). Retrieved from <http://www.eesc.europa.eu/?i=portal.en.theme-europe-2020-flagship-initiatives-digital-agenda-opinions>

European Parliament. (2015). | Search Results | European Parliamentary Research Service Blog. Retrieved 28 September 2016, from <https://epthinktank.eu/?s=>

European Parliament and the Council of the European Union. (2009, November 25). DIRECTIVE 2009/136/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 25 November 2009 amending Directive 2002/22/EC on universal service and users' rights relating to electronic communications networks and services, Directive 2002/58/EC concerning the processing of personal data and the protection of privacy in the electronic communications sector and Regulation (EC) No 2006/2004 on cooperation between national authorities responsible for the enforcement of consumer protection laws. Retrieved from <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:337:0011:0036:en:PDF>

Eurostat. (2014). *Eurostat: employment and unemployment LFS*.

EUROSTAT. (2016a). Digital single market - promoting e-commerce for individuals - Eurostat. Retrieved 20 September 2016, from http://ec.europa.eu/eurostat/web/products-datasets/-/isoc_bdek_smi

EUROSTAT. (2016b). Enterprises selling via internet and/or networks other than internet (NACE Rev. 2 activity) - Eurostat. Retrieved 20 September 2016, from http://ec.europa.eu/eurostat/web/products-datasets/-/isoc_ec_eseln2

EUROSTAT. (2016c). GDP and main components (output, expenditure and income) - Eurostat. Retrieved 20 September 2016, from http://ec.europa.eu/eurostat/web/products-datasets/-/nama_10_gdp

Eurostat. (2016). Internet purchases by individuals.

Extra, Guus, & Kutlay, Yağmur. (2012). Trends in policies and practices for multilingualism in Europe. British Council. Retrieved from https://englishagenda.britishcouncil.org/sites/default/files/attachments/lre_english_version_final_01.pdf

Feenstra, R. C. (2002). Border effects and the gravity equation: consistent methods for estimation. *Scottish Journal of Political Economy*, 49(5), 491–506.

Feinerer, I., & Hornik, K. (2012). tm: Text mining package. *R Package Version 0.5-7.1*, 1(8).

Fidrmuc, J. (2011). The Economics of Multilingualism in the EU. *Economics and Finance Working Paper Series, Working Paper*, (4).

Foreman-Peck, J., & Wang, Y. (2013a). *The costs to the UK of Language Deficiencies as a Barrier to UK Engagement in Exporting*. UK Trade & Investment; Cardiff Business School. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/309899/Costs_to_UK_of_language_deficiencies_as_barrier_to_UK_engagement_in_exporting.pdf

Foreman-Peck, J., & Wang, Y. (2013b). *The costs to the UK of Language Deficiencies as a Barrier to UK Engagement in Exporting*. UK Trade & Investment; Cardiff Business School. Retrieved from

- https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/309899/Costs_to_UK_of_language_deficiencies_as_barrier_to_UK_engagement_in_exporting.pdf
- Galili, T. (2015). dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btv428>
- Gazzola, M. (2014). Language Policy and Linguistic Justice in the European Union: The Socio-Economic Effects of Multilingualism.
- Gazzola, M. (2016). Multilingual communication for whom? Language policy and fairness in the European Union. *European Union Politics*, 1465116516657672.
- Gazzola, M., & Grin, F. (2013). Is ELF more effective and fair than translation? An evaluation of the EU's multilingual regime. *International Journal of Applied Linguistics*, 23(1).
- Graves-Brown, P., Jones, S., & Gamble, C. S. (2013). *Cultural identity and archaeology: the construction of European communities*. Routledge.
- Hagen, S., Foreman-Peck, J., Davila-Philippson, S., Nordgren, B., & Hagen, S. (2006). ELAN: Effects on the European economy of shortages of foreign language skills in enterprise. Brussels: European Commission. Http://Ec.Europa.Eu/Languages/Languages-Mean-Business/Files/Elan-Full-Report_en.Pdf.
- Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., & Morris, M. (2008). statnet: Software tools for the representation, visualization, analysis and simulation of network data. *Journal of Statistical Software*, 24(1), 1548.
- Hlavac, M. (2015). *stargazer: Well-Formatted Regression and Summary Statistics Tables*. Cambridge, USA: Harvard University. Retrieved from <http://CRAN.R-project.org/package=stargazer>
- Højsgaard, S., & Halekoh, U. (2016). *doBy: Groupwise Statistics, LSmeans, Linear Contrasts, Utilities*. Retrieved from <https://CRAN.R-project.org/package=doBy>
- Hornik, K. (2016). *NLP: Natural Language Processing Infrastructure*. Retrieved from <https://CRAN.R-project.org/package=NLP>
- Hunter, J., & Wilson, M. (2015, December 2). CROSS-BORDER ONLINE SHOPPING WITHIN THE EU. LEARNING FROM CONSUMER EXPERIENCES. ANEC, the European consumer voice in standardisation. Retrieved from <http://www.anec.eu/attachments/ANEC-RT-2015-SERV-005.pdf>
- India Government. (NA). About ILDC. Retrieved 27 October 2016, from <http://ildc.in/>
- India Government. (2005). *Right To Information Act 2005*. Retrieved from <http://odishapolice.gov.in/sites/default/files/PDF/Right%20To%20Information%20Act%202005.pdf>
- Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing* (2nd ed.). Prentice Hall. Retrieved from <https://web.stanford.edu/~jurafsky/slp3/>
- Kahle, D., & Wickham, H. (2013). ggmap: Spatial Visualization with ggplot2. *The R Journal*, 5(1), 144–161.
- Kastberg, C., Jervelund, C., Rytz, B., Hansen, S., Ramskov, J., & Gros, K. (2007). *Study on The Economic Impact of the Electronic Commerce Directive*. DG Internal Market and Services, European Commission. Retrieved from http://ec.europa.eu/internal_market/e-commerce/docs/study/ecd/%20final%20report_070907.pdf
- Kleiber, C., & Zeileis, A. (2008). *Applied Econometrics with R*. New York: Springer-Verlag. Retrieved from <http://CRAN.R-project.org/package=AER>

- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., ... Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL-2007 Demo and Poster Sessions*. Prague, Czech Republic.
- Lahti, L., Huovari, J., Kainu, M., & Biecek, P. (2014). *eurostat R package*. Retrieved from <https://github.com/rOpenGov/eurostat>
- Landan, D. (2016, May). Neural Machine Translation is the Next Big Thing. Retrieved from <https://www.welocalize.com/neural-machine-translation-next-big-thing/>
- LRMA. (2011). Aims of the RMA. Retrieved 27 October 2016, from <http://rma.nwu.ac.za/index.php//aims/>
- LT-Innovate. (2012a). *Establishing Europe's Global Market Position & Securing the Digital Single Market*. LT-Innovate. Retrieved from http://www.lt-innovate.org/sites/default/files/documents/2074-LTIA_Report_20130214_Med_Res.pdf
- LT-Innovate. (2012b, December 6). *Establishing Europe's Global Market Position & Securing the Digital Single Market. An industry vision*. Retrieved from <http://www.eesc.europa.eu/resources/docs/digitalmarketppen.pdf>
- LT-Innovate. (2013a). *Status and Potential of the European Language Technology Markets*. Retrieved from http://www.lt-innovate.org/sites/default/files/2216-LT2013_Report_MediumQuality.pdf#overlay-context=lt-observe/document/lt-innovate-innovation-agenda-manifesto
- LT-Innovate. (2013b). *Status and Potential of the European Language Technology Markets*. Retrieved from http://www.lt-innovate.org/sites/default/files/2216-LT2013_Report_MediumQuality.pdf#overlay-context=lt-observe/document/lt-innovate-innovation-agenda-manifesto
- LT-Innovate. (2015, June). *European Platform for the Multilingual Digital Single Market*. Retrieved from <http://www.lt-innovate.org/sites/default/files/Multilingual%20Platform%20Concept.pdf#overlay-context=elc>
- LT-Innovate. (2016a). *The LT-Innovate Innovation Agenda*. Retrieved from http://www.lt-innovate.org/sites/default/files/2904-LTi_Innovation_Agenda.pdf
- LT-Innovate. (2016b). *The LT-Innovate Innovation Manifesto*. Retrieved from http://www.lt-innovate.org/sites/default/files/2904-LTI_Manifesto_201406_SmallFileSize.pdf#overlay-context=lt-observe/resources/public-positions
- LT-Innovate. (2016c, November). *Response to European Parliament /STOA Experts' questionnaire on Market and Economic Impact of the Human Language Technology Sector*. Retrieved from <http://www.lt-innovate.org/sites/default/files/STOA%20HLT%20Interview%20-%20LT-Innovate.pdf>
- M. Prys Jones. (2013). *Endangered languages and linguistic diversity in the European Union*. European Parliament. Retrieved from [http://www.europarl.europa.eu/RegData/etudes/note/join/2013/495851/IPOL-CULT_NT\(2013\)495851\(SUM01\)_EN.pdf](http://www.europarl.europa.eu/RegData/etudes/note/join/2013/495851/IPOL-CULT_NT(2013)495851(SUM01)_EN.pdf)
- Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press. Retrieved from <http://nlp.stanford.edu/fsnlp/>
- META technology council. (2012). *Strategic Research Agenda for Multilingual Europe 2020*. Germany: META-NET. Retrieved from http://www.meta-net.eu/vision/reports/meta-net-sra-version_1.0.pdf
- Meyer, D., Hornik, K., & Feinerer, I. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, 25(5), 1-54.

- Muller, P., Caliandro, C., Peycheva, V., Gagliardi, D., Marzocchi, C., Ramlogan, R., & Cox, D. (2014). Annual Report on European SMEs. *Final Report-July*. Retrieved from <http://www.pmievolution.it/wp-content/uploads/2016/04/annual-report-SME-2015.pdf>
- Nations Online Project. (2016). European Languages by Countries. Official and national Languages of Europe. Nations Online Project. Retrieved 5 November 2016, from http://www.nationsonline.org/oneworld/european_languages.htm
- NPLD. (2016). Network to Promote Linguistic Diversity. Retrieved 14 November 2016, from <http://www.npld.eu/>
- NPLD, & EFNIL. (2016, July). Joint Position Paper NPLD - EFNIL 2016 Joint NPLD - EFNIL Position paper on Language and Technology Use the technology or lose your language? Will new communication tools and welfare robots speak our languages! Retrieved from <http://www.npld.eu/uploads/publications/356.pdf>
- Pavel, F. (2010, 03). A Single Market for an Information Society – Economic Analysis Final Report for the Directorate General Information Society and Media. DIW econ GmbH. Retrieved from http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?doc_id=887
- Quan, K., & Lynch, J. (2010). *The High Costs of Language Barriers in Medical Malpractice*. School of Public Health - University of California, Berkley. Retrieved from http://www.pacificinterpreters.com/docs/resources/high-costs-of-language-barriers-in-malpractice_nhelp.pdf
- QUARTZ. (2016, 9). Google Translate (GOOG) now uses artificial intelligence and scores nearly identically to human translators. Retrieved 16 November 2016, from <http://qz.com/792621/googles-new-ai-powered-translation-tool-is-nearly-as-good-as-a-human-translator/>
- R Core Team. (2016a). *foreign: Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, ...* Retrieved from <https://CRAN.R-project.org/package=foreign>
- R Core Team. (2016b). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- Rehm, G. (2017, January). *Human Language Technologies in a Multilingual Europe*. Presented at the STOA Workshop on Language Technologies, EP, Brussels.
- Rehm, G., & Uszkoreit, H. (2012). *Strategic Research Agenda for Multilingual Europe 2020*. Germany: META-NET Technology Council. Retrieved from http://www.meta-net.eu/vision/reports/meta-net-sra-version_1.0.pdf
- Rehm, G., Uszkoreit, H., Ananiadou, S., Bel, N., Bielevičienė, A., Borin, L., ... Daelemans, W. (2016). The strategic impact of META-NET on the regional, national and international level. *Language Resources and Evaluation*, 1-24.
- Rehm, G., Uszkoreit, H., Dagan, I., Goetcherian, V., Dogan, M. U., Mermer, C., ... Prys Jones, M. (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 OpenData Era*.
- Renals, S., Carletta, J., Edwards, K., Bourlard, H., Garner, P., Popescu-Belis, A., ... Wacker, P. (2015). *Roadmap for Conversational Interaction Technologies*. Retrieved from http://www.lt-innovate.org/sites/default/files/citia_files/rockit-scenarios-whitepaper-v2.1.pdf#overlay-context=citia/roadmap
- Rinker, T. W. (2013). *qdap: Quantitative Discourse Analysis Package*. Buffalo, New York: University at Buffalo/SUNY. Retrieved from <http://github.com/trinker/qdap>

- ROCKIT Project: Roadmap for Conversational Interaction Technologies. (n.d.). Retrieved from <http://rockit-project.eu/>
- Rossi, K. (2016). *Horizon 2020 and Language Technology (LT)*. Brussels. Retrieved from http://www.lt-innovate.org/sites/default/files/lt_summit/09.30%20Kimmo%20Rossi.pdf
- Rudis, B. (2016). *ggalt: Extra Coordinate Systems, Geoms and Statistical Transformations for 'ggplot2'*. Retrieved from <https://CRAN.R-project.org/package=ggalt>
- SETSI. (2015). *Plan de Impulso de Tecnologías del Lenguaje*. Retrieved from <http://www.agendadigital.gob.es/planes-actuaciones/tecnologias-lenguaje/Bibliotecaimpulsotecnologiaslenguaje/Detalle%20del%20Plan/Plan-impulso-Tecnologias-Lenguaje.pdf>
- Silva, J. S., & Tenreyro, S. (2006). The log of gravity. *The Review of Economics and Statistics*, 88(4), 641–658.
- Silva, J. S., & Tenreyro, S. (2011). Further simulation evidence on the performance of the Poisson pseudo-maximum likelihood estimator. *Economics Letters*, 112(2), 220–222.
- SIRIA. (2016). *Strategic Research and Innovation Agenda. Language as a Data Type and Key Challenge for Big Data*. Retrieved from <http://www.cracking-the-language-barrier.eu/wp-content/uploads/SRIA-V0.9-final-online.pdf>
- Skutnabb-Kangas, T. (2002). *Why should linguistic diversity be maintained and supported in Europe? Some arguments*. Strasbourg: Council of Europe. Retrieved from <https://www.coe.int/t/dg4/linguistic/Source/Skutnabb-KangasEN.pdf>
- Slowikowski, K. (2016). *ggrepel: Repulsive Text and Label Geoms for 'ggplot2'*. Retrieved from <https://CRAN.R-project.org/package=ggrepel>
- South Africa Government. (NA). *hlt_booklet.pdf*. Retrieved 27 October 2016, from http://www.csir.co.za/meraka/hlt/docs/hlt_booklet.pdf
- South Africa Government. (2003). *LPD Language Policy*. Retrieved from http://www.dac.gov.za/sites/default/files/LPD_Language%20Policy%20Framework_English_0.pdf
- Spanish Government. (2013). *Digital Agenda for Spain*. Retrieved 27 October 2016, from <http://www.agendadigital.gob.es/digital-agenda/Paginas/digital-agenda-spain.aspx>
- The Guardian. (2014, January 27). Google buys UK artificial intelligence startup Deepmind for £400m | Technology | The Guardian. Retrieved 2 February 2017, from <https://www.theguardian.com/technology/2014/jan/27/google-acquires-uk-artificial-intelligence-startup-deepmind>
- TNS. (2015, November 30). European Parliament Eurobarometer (EB/EP 84.1). ANALYTICAL OVERVIEW. Retrieved 27 October 2016, from http://www.europarl.europa.eu/pdf/eurobarometre/2015/2015parlemeter/eb84_1_synthese_analytique_partie_II_en.pdf
- UN. (2016). *International Migration Report 2015*. New York. Retrieved from http://www.un.org/en/development/desa/population/migration/publications/migrationreport/docs/MigrationReport2015_Highlights.pdf
- USCB. (2015, November). *Geographical Mobility: 2014 to 2015*. Retrieved 27 October 2016, from <http://www.census.gov/data/tables/2015/demo/geographic-mobility/cps-2015.html>
- Vanian, J. (2016, July 12). *Why Data Is The New Oil*. Retrieved from <http://fortune.com/2016/07/11/data-oil-brainstorm-tech/>

- Vries, A. de, & Ripley, B. D. (2016). *ggdendro: Create Dendrograms and Tree Diagrams Using 'ggplot2'*. Retrieved from <https://CRAN.R-project.org/package=ggdendro>
- Wang, J., Zamar, R., Marazzi, A., Yohai, V., Salibian-Barrera, M., Maronna, R., ... Konis, K. (2014). *robust: Robust Library*. Retrieved from <https://CRAN.R-project.org/package=robust>
- Weidmann, N. B., & Gleditsch, K. S. (2015). *cshapes: CShapes Dataset and Utilities*. Retrieved from <https://CRAN.R-project.org/package=cshapes>
- Weidmann, N. B., Kuse, D., & Gleditsch, K. S. (2011). *cshapes: CShapes Dataset and Utilities. R Package Version 0.3-1*, URL <Http://CRAN.R-Project.Org/Package=Cshapes>.
- Weiß, J., & Schwietring, T. (NA). *The Power of Language A philosophical-sociological reflection*.
- Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. Retrieved from <http://ggplot2.org>
- Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software*, 40(1), 1–29.
- Wickham, H. (2016a). *rvest: Easily Harvest (Scrape) Web Pages*. Retrieved from <https://CRAN.R-project.org/package=rvest>
- Wickham, H. (2016b). *scales: Scale Functions for Visualization*. Retrieved from <https://CRAN.R-project.org/package=scales>
- Wickham, & Hadley. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12). Retrieved from <http://www.jstatsoft.org/v21/i12/paper>
- Wikipedia. (2016). Czech Republic - Wikipedia. Retrieved 5 November 2016, from https://en.wikipedia.org/wiki/Czech_Republic#cite_note-3
- World Bank. (2016). Worldwide Governance Indicators. Retrieved 10 October 2016, from <http://info.worldbank.org/governance/wgi/index.aspx#home>
- YOTPO. (2015). The State of eCommerce: Yotpo Benchmark Report. Retrieved from http://cdn2.hubspot.net/hubfs/462213/Yotpo_eBook_Benchmark_2015.pdf
- Zeileis, A. (2004). Econometric computing with HC and HAC covariance matrix estimators.
- Zeileis, A. (2006). Object-oriented computation of sandwich estimators.
- Zeileis, A., & Hothorn, T. (2002). Diagnostic Checking in Regression Relationships. *R News*, 2(3), 7–10.

7 Annexes

7.1 List of experts interviewed and acknowledgments

The following experts were interviewed during the project. We want to thank them for their valuable ideas and contributions:

- Nuria Bel, Professor at the Department of Translation and Language Sciences and senior researcher at the Applied Linguistics Institute of the Universitat Pompeu Fabra
- Gerhard Budin, Professor at the Centre of Translation Studies at the University of Vienna. Director of the Institute for Corpus Linguistics and Text Technology at the Austrian Academy of Sciences. Chair holder of the UNESCO Chair for Multilingual, Transcultural Communication in the Digital Age.
- Stanislas Dehaene, Professor at the Collège de France and Neuroscience Specialist.
- Michael Fritz, Executive Director at European Association for Technical Communication - tekomp Europe e.V.
- John Judge, EU Research Manager at ADAPT Centre.
- Rose Lockwood, Director of Market Research at LT-Innovate and trend watcher at TAUS.
- Joseph Mariani, Director of Research at CNRS-IMMI & LIMSI.
- Adriane Rinsche, Managing Director at LTC Innovates.
- Peggy van der Kreeft, Innovation Manager, New Media, at Deutsche Welle.
- Josef van Genabith, Head of the Multilingual Technologies group at DFKI GmbH.
- Philippe Wacker, Secretary-General at LT-Innovate.
-

The following experts and organizations reviewed the draft version of the study and provided valuable feedback. We also want to thank them for their contribution to the project:

- The European Language Equality Network (ELEN)
- The Digital Language Diversity Project (DLDP)
- Khalid Choukri, European Language Resources Association (ELRA) Secretary General; ELDA founder & CEO
- Horváth István, President, Romanian Institute for Researching the Problems of National Minorities
- Sabine Kirchmeier, Vice-President of the executive committee, European Federation of National Institutions of Languages
- Marco Marsella, Head of Unit, European Commission DG Connect Unit G3 – Learning, Multilingualism & Accessibility
- Georg Rehm, Senior Researcher, German Research Centre for Artificial Intelligence and General Secretary, Network of Excellence META-NET
- Hans Uszkoreit, Scientific Director, German Research Centre for Artificial Intelligence and Network of Excellence META-NET
- Andrejs Vasiljevs, CEO, Tilde
- Aleksandra Wesolowska, Programme Officer – EU policies, European Commission DG CONNECT Unit G3 – Learning, Multilingualism & Accessibility

We would also like to thank the R Development Core Team for providing the tool used in the quantitative analysis (R Core Team, 2016b), the developers of the R-studio environment (Studio, 2012) and all the developers of the R libraries used in the analysis (Auguie, 2016; Bivand & Lewin-Koh, 2016; Bojanowski, 2015; Brownrigg, Minka, & Bivand, 2015; Csardi & Nepusz, 2006; Feinerer & Hornik, 2012; Galili, 2015; Handcock, Hunter, Butts, Goodreau, & Morris, 2008; Hlavac, 2015; Højsgaard & Halekoh, 2016; Hornik, 2016; Kahle & Wickham, 2013; Kleiber & Zeileis, 2008; Lahti, Huovari, Kainu, & Biecek, 2014; Meyer, Hornik, & Feinerer, 2008; R Core Team, 2016a; Rinker, 2013; Rudis, 2016; Slowikowski, 2016; Vries & Ripley, 2016; Wang et al., 2014; Nils B. Weidmann & Gleditsch, 2015; Wickham, 2009, 2011, 2016a, 2016b; Wickham & Hadley, 2007; Zeileis, 2004, 2006; Zeileis & Hothorn, 2002). They made quantitative analysis feasible.

7.2 List of official languages by country

The list of official languages is obtained from (Nations Online Project, 2016). We make an exception in Czech Republic where Slovak may be considered an official language (Wikipedia, 2016).

Table 19: Official languages by country

Country	Official languages
Austria	German, Slovenian, Croatian, Hungarian
Belgium	Dutch, French, German
Bulgaria	Bulgarian
Croatia	Croatian
Cyprus	Greek, Turkish, English
Czech Republic	Czech, Slovak
Denmark	Danish
Estonia	Estonian
Finland	Finnish, Swedish
France	French
Germany	German
Greece	Greek
Hungary	Hungarian
Ireland	Irish, English
Italy	Italian
Latvia	Latvian
Lithuania	Lithuanian
Luxembourg	Luxembourgish, French, German
Malta	Maltese
Netherlands	Dutch
Poland	Polish
Portugal	Portuguese
Romania	Romanian
Slovakia	Slovak

Country	Official languages
Slovenia	Slovenian
Spain	Spanish, Catalan, Galician, Basque
Sweden	Swedish
United Kingdom	English

Source: Compiled by the authors based in (Nations Online Project, 2016; Wikipedia, 2016)

7.3 Estimation of the effect of socio-demographic factors on speaking a foreign language

The analysis is based on the dataset of the Eurobarometer 77.1 that include a topic about Multilingualism (European Commission, 2014a). There are specific questions in the survey about the foreign languages spoken and the level of knowledge of the language (very good, good, and basic) in 2011 along with socio-demographic characteristics of the respondents. The effect of the socio-demographic factor on the chance of speaking a foreign language (at least good) is assessed by using a Logistic regression. The response variable is whether or not the individual speaks, with a level at least good, the languages under analysis (excluding the population that have any of the analysed languages as their mother tongue). The results of the odds ratio (the increased chance of speaking the selected languages for each characteristic) are shown in Table 20.

Table 20: LOGIT Regression results

Weighted LOGIT Regression results for different languages depending on socio-demographic characteristics (Odds Ratios) excluding those whose mother tongue is one of the analysed

	<i>Dependent variable:</i>	
	English language	6 majority languages
age: 50 to 64	1.630 t = 4.711***	1.501 t = 4.426***
age: 30 to 49	2.955 t = 10.995***	2.499 t = 9.715***
age: below 30	5.101 t = 14.126***	4.271 t = 11.021***
education: Secondary	4.762 t = 11.506***	2.462 t = 7.299***
education: University	19.205 t = 21.517***	8.998 t = 17.254***
social status: Medium	2.043 t = 5.321***	1.569 t = 3.363***
social status: High	3.710 t = 8.312***	3.106 t = 7.184***
place: Small/middle town	1.237 t = 2.867***	1.615 t = 6.287***
place: Large town	1.443 t = 4.647***	2.045 t = 8.517***
gender: Male	1.077 t = 1.239	1.033 t = 0.519
Observations	21,599	15,459

Note:

*p<0.1; **p<0.05; ***p<0.01

Base case: Age 65 and above; Primary education; Low social status; Living in Rural area or village; Female

Source: Compiled by the authors based in the Eurobarometer 77.1 (European Commission, 2014a)

7.4 Cluster analysis of cross-border e-commerce

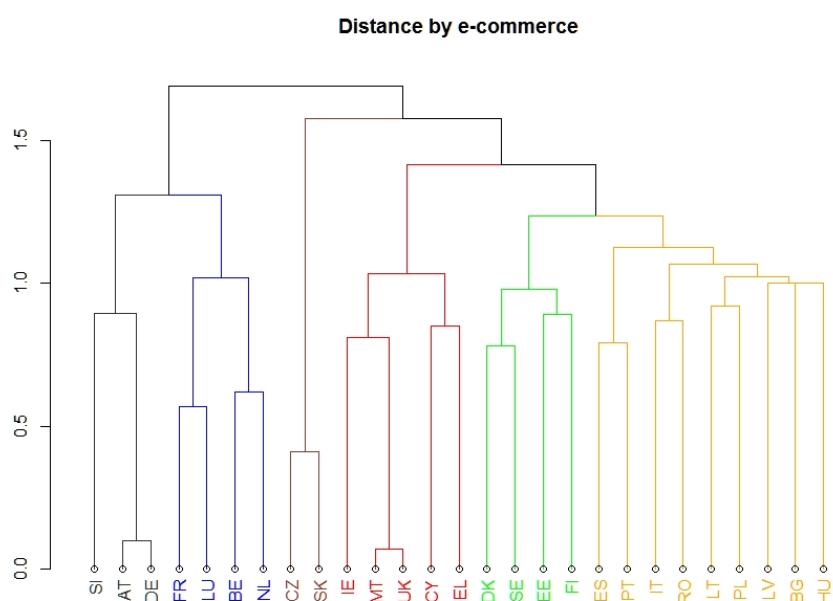
7.4.1 Demand side

The percentage of on-line shoppers from one country buying in another country in 2011⁴¹ out of the total of on-line shoppers is included in the study of Civic Consulting (2011). Based on this figure for each pair of countries, the distance between two countries is: 1-(the highest of the percentage of on-line shoppers from one country buying in the other country/100). Those pairs of countries which a higher cross-border of e-commerce are closer.

⁴¹ We have seen that the cross-border e-commerce share in total e-commerce has not substantially changed since 2011 so the estimates are likely to be valid in 2016.

The analysis is made by using the Ward algorithm. The clusters as shown in the dendrogram in Figure 45.

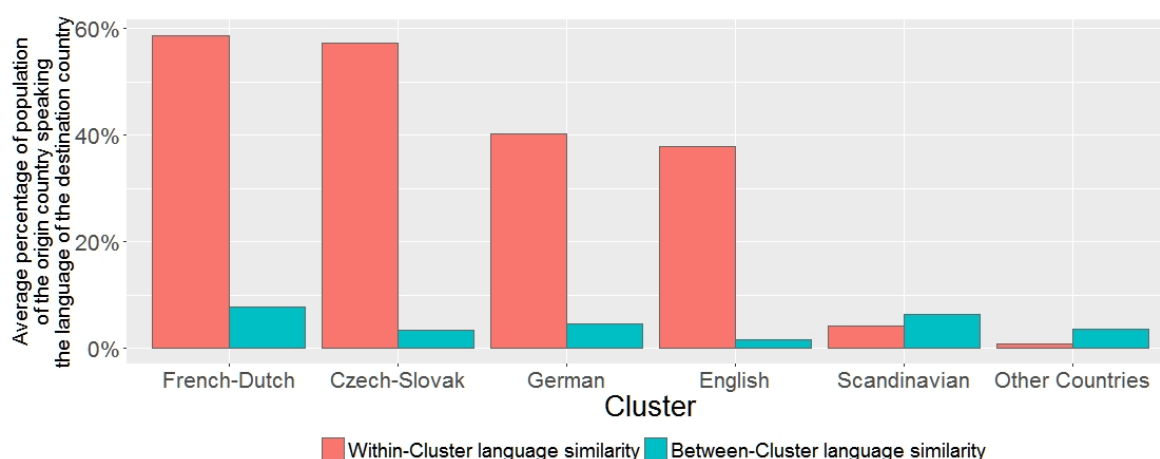
Figure 45: Dendrogram of clusters of cross-border e-commerce (consumers' side)



Source: Compiled by the authors based on (Civic Consulting, 2011)

The clusters are usually grouped around a big country while most of the countries within the clusters share the same or very similar languages (particularly the cluster that includes the rest of EU countries that have different languages (we called it the “Other Countries” cluster) as shown in Figure 46.

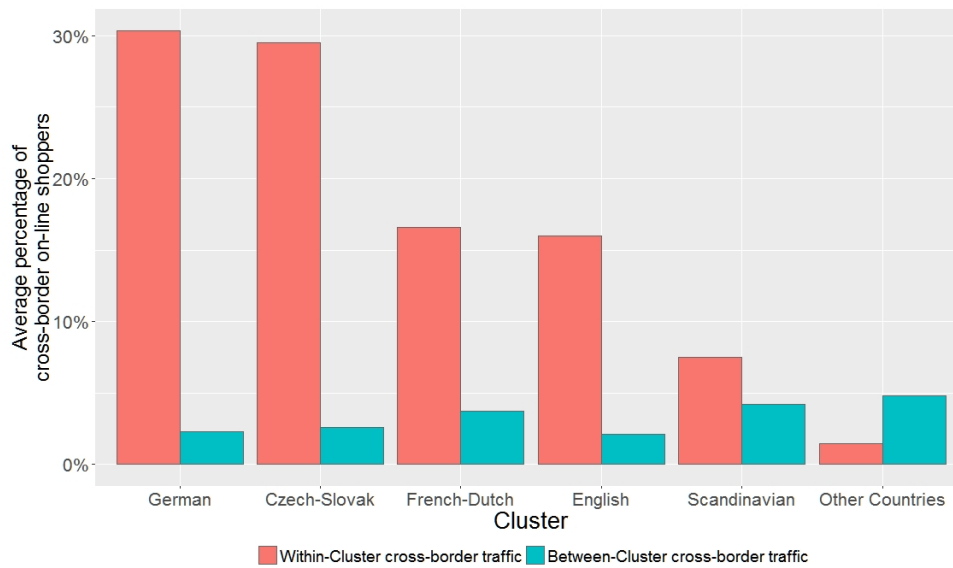
Figure 46: Average percentage of population of the origin country speaking the language of the destination country by cluster (consumers' side)



Source: Compiled by the authors based on (Civic Consulting, 2011; European Commission, 2014a; EUROSTAT, 2016c)

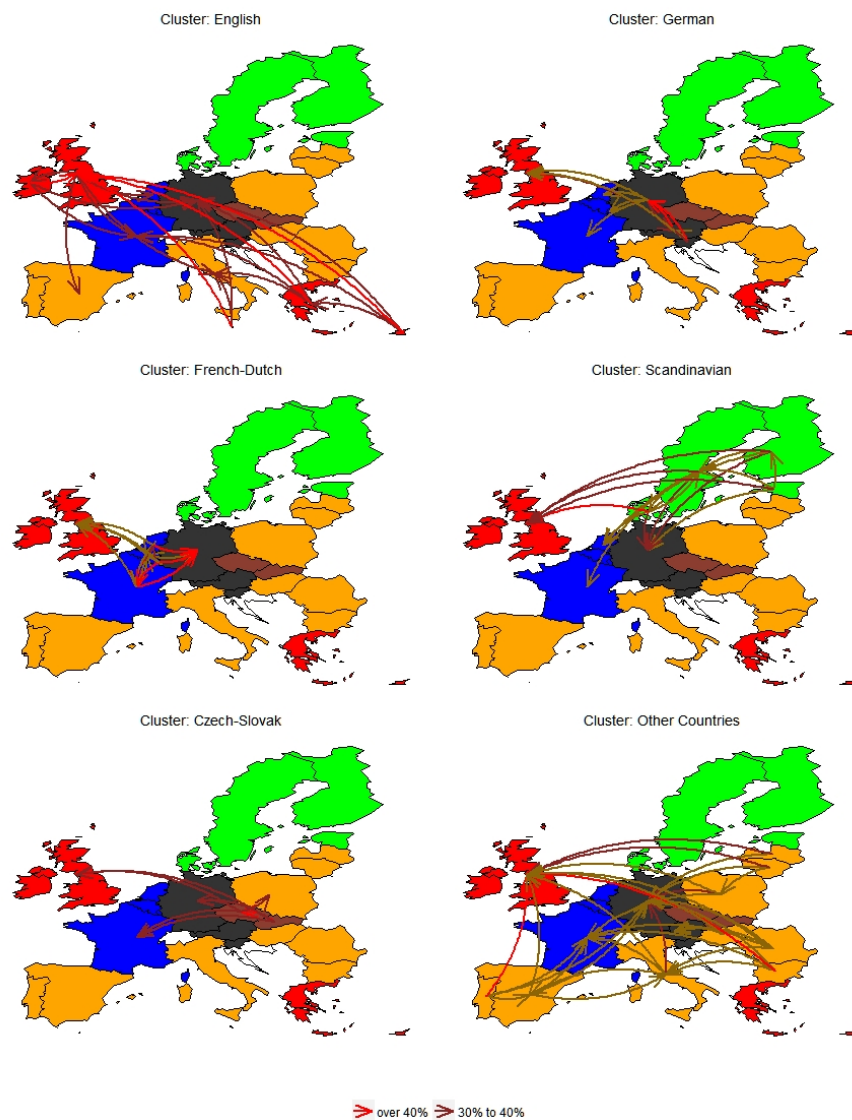
The clusters share similar patterns of cross-border e-commerce. The three clusters around Germany, United Kingdom and France along with the Czech-Slovak cluster show much higher cross-border traffic within the countries of the cluster compared to cross-border traffic with countries of other clusters. On the other hand, in the “Other Countries” cluster, that includes disparate countries with different languages, the opposite applies.

Figure 47: Comparison of within and between average cross-border e-commerce by cluster (consumers' side)



Source: Compiled by the authors based on (Civic Consulting, 2011; European Commission, 2014a; EUROSTAT, 2016c)

Regardless of the cluster, most of the e-shoppers have also strong e-commerce ties with the UK and Germany and to some extent with France, the champions of European cross-border e-commerce as seen in Figure 48. The scarce cross-border traffic between countries of different clusters is mainly captured by those big countries.

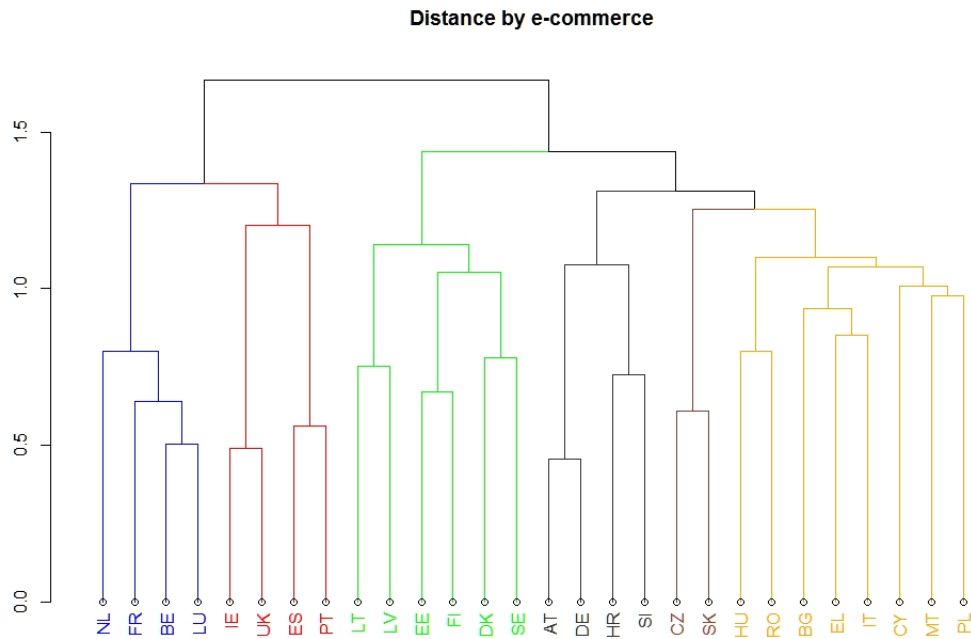
Figure 48: Cross-border e-commerce by cluster (consumers' side)

Source: Compiled by the authors based on (Civic Consulting, 2011)

7.4.2 Supply side

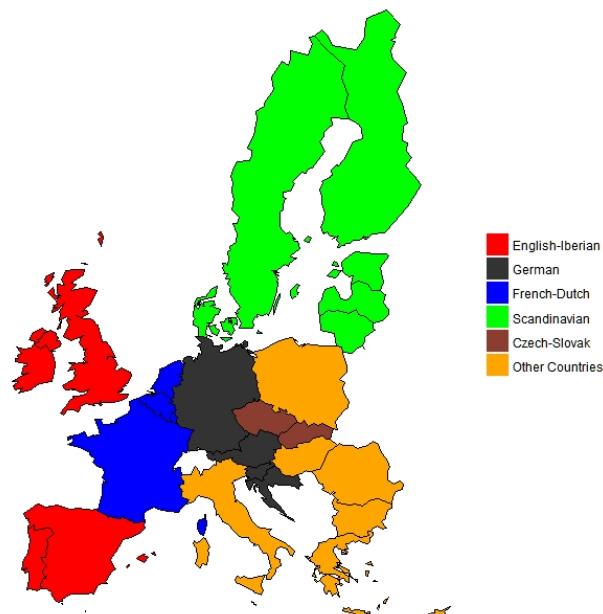
The variable to estimate the distance between countries is the number of on-line cross-border retailers selling to individual consumers in 2015. To calculate this value we use the dataset of the Flash Eurobarometer 413 – Companies Engaged in Online Activities – (European Commission, 2015c), that includes information about the countries where the retailer is selling to. Based on this figure for each pair of countries, the distance between two countries is: $1 - (\text{the highest of the percentage of on-line retailers from one country selling to the other country}) / 100$.

The analysis is made by using the Ward algorithm. We find a similar pattern of 6 clusters as shown in the dendrogram in Figure 49.

Figure 49: Dendrogram of clusters of cross-border e-commerce (retailers' side)

Source: Compiled by the authors based on (European Commission, 2015c)

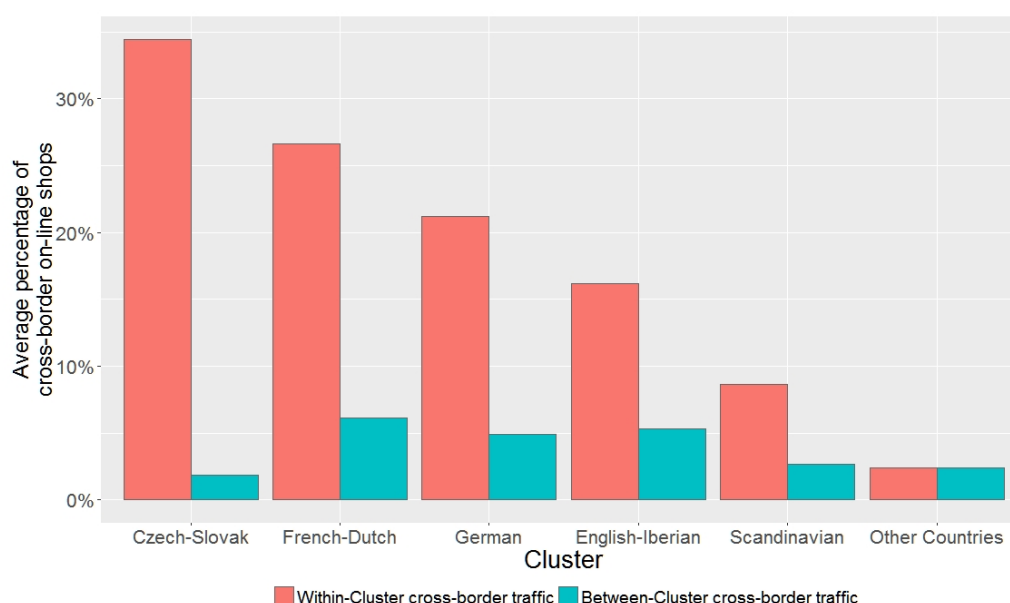
The main difference is that Spain and Portugal become part of the called English-Iberian group while Greece becomes a member of the “Other Countries” group. There are also some minor differences regarding the Baltic republics that become part of the Scandinavian group. The clusters are still mainly grouped around a big country while most of the countries within the clusters share the same or very similar languages as shown in Figure 50.

Figure 50: Clusters of cross-border e-commerce (retailers' side)

Source: Compiled by the authors based on (European Commission, 2015c)

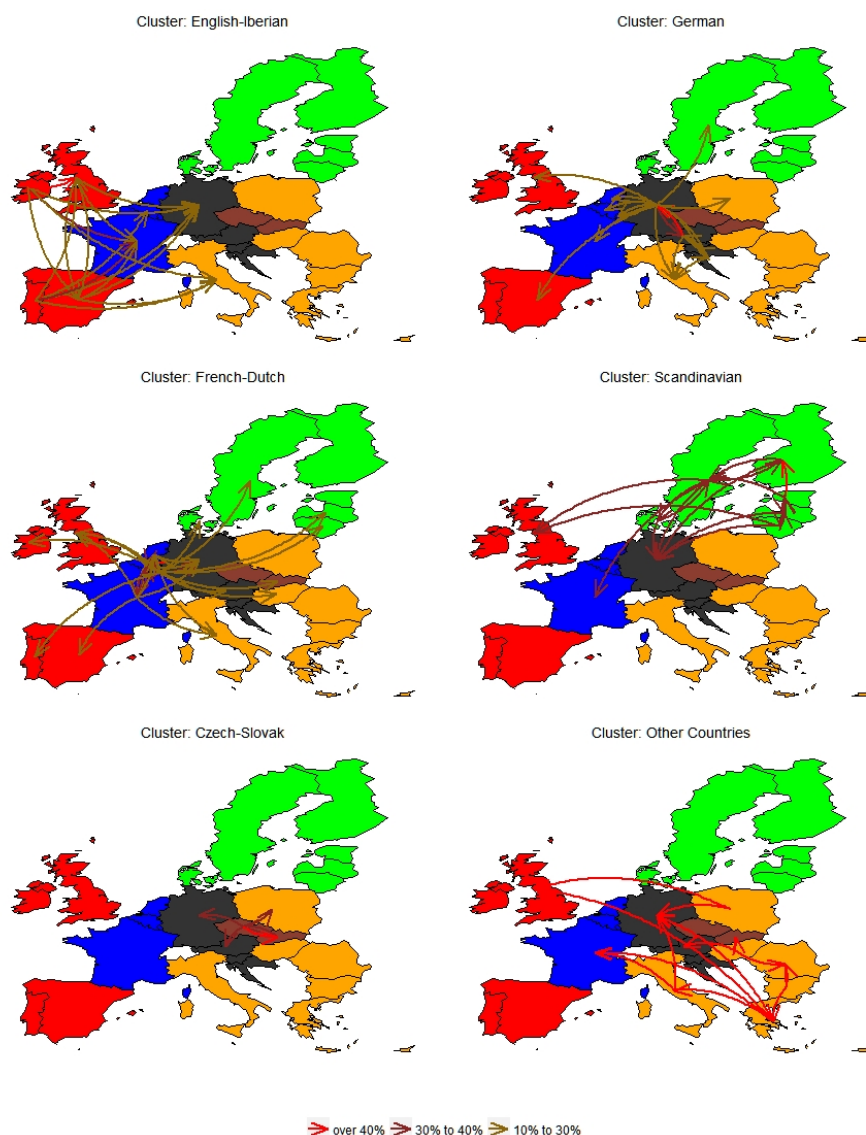
The retailers usually sell to countries within the same group, except the “Other Countries” group. It is particularly interesting to analyse this group. E-sellers in those countries do not tend to sell to other countries, neither within the group nor between groups. In fact, the percentage of companies selling on-line to other countries is the lowest. It makes sense if considering that countries in that group do not share any language and therefore it is particularly challenging for them to sell to other countries. It is particularly concerning because most of the countries of that cluster are smaller language countries that become specially disadvantaged due to the unattended language barriers.

Figure 51: Comparison of within and between average cross-border e-commerce by cluster (retailers’ side)



Source: Compiled by the authors based on (Civic Consulting, 2011; European Commission, 2014a; EUROSTAT, 2016c)

These clusters also share similar patterns of e-commerce as seen in Figure 52. Regardless of the cluster, most of the retailers tend to sell to the big countries such as UK, Germany and France, the biggest economies and therefore more potentially interesting markets. The scarce cross-border offering between countries of different clusters is mainly directed to those big countries. The main conclusion is that on-line shop retailers in Europe is strongly fragmented into six groups mainly shaped by language similarities and the retailers do not tend to sell to other countries out of their clusters except to the big economies. Eventually, the countries that do not share any language (mainly in the “Other Countries” cluster) become particularly disadvantaged.

Figure 52: Cross-border e-commerce by cluster (retailers' side)

Source: Compiled by the authors based on (European Commission, 2015c)

7.5 Estimation of the effect of language barriers using regression analysis

We analyse the effect of language barriers by using the "gravity model" of cross-border international trade that is the workhorse for analysing international flows. The "gravity model" is equivalent to the Newton's gravity theory where the force of attraction (volume of international flows) is related to the size of the planets (countries) and the distance between them. The distance can capture, not only the physical distance but other components of the relationship between the countries, such as language "distance", or specific features of the origin and destination countries that could affect the flows (such as regulatory quality, unemployment rate, etc.). A detailed explanation of the gravity model applied to e-commerce can be found in (Cardona et al., 2015).

In this study we analyse the effect of language barriers in three variables, namely cross-border workers' mobility, cross-border e-commerce (demand side), and cross-border e-commerce (supply side).

7.5.1 Analysis of the cross-border workers' mobility model

We analyse the flow of workers' mobility between the EU27 countries in 2009. Croatia is excluded of the analysis because there were no data available of this country in the Eurobarometer 77.1 (European Commission, 2014a). We create a dataset of 702 edges (27 origin countries x 27 destination countries excluding the pairs of a country to itself).

7.5.1.1 Response variable

The response variable is the EU population that is currently living and working in a country different to its citizenship or that have do so in the past. We use the Eurobarometer 72.5 dataset (European Commission, 2009) that includes a specific topic about geographical and labour market mobility. If the citizen is currently living and working in a country different to its citizenship that is the observation that we include in the model. The destination country is the country where he is currently living. The dataset also includes information about the last country where a citizen lived and worked. If the citizen is currently living in its country and the citizen has mentioned that he has lived and worked in a foreign country in the past that is the observation we include in the model. The destination country is the last foreign country where he lived and worked. There are 12 observations of citizens with dual nationality out of 1 842 observations. We include those 12 observations as different citizens. Another strategy would have been to drop those observations out of the model. Whatever the strategy we follow the results will almost be the same due to the low number of observations.

7.5.1.2 Independent variable

To calculate the language barriers between the countries we use the dataset of the Eurobarometer 77.1 that include a topic about Multilingualism (European Commission, 2014a). There are specific questions in the survey about the foreign languages spoken and the level of knowledge of the language (very good, good, and basic) in 2011. Based on that questions we classify the language similarities between the origin country and the destination country into four groups, namely low language barriers (more than 85 % of the population in the origin country speaks, at least good, one of the majority official languages of the destination country⁴²), medium-low language barriers (between 50 % and 85 %), medium-high language barriers (between 15 % and 50 %) and high language barriers (less than 15 %). We are using data of 2011 because this is the only available dataset that provides rich information about foreign languages. Although using data of 2011 when analysing the effect on a variable of 2009 may introduce some error we expect that the results are still reliable because we have seen that the changes in the percentage of speakers of a foreign language in a country are very likely to evolve very slowly (see Figure 15).

7.5.1.3 Control variables

The market size of the destination country is the GDP of the country in 2009⁴³ (EUROSTAT, 2016c).

The size of the origin country is the total population of the country in 2009 (European Commission, 2016g).

⁴² We consider that the official national or regional language is a majority language in the country if it is spoken, with a level at least good, by more than 10 % of the population. Using this procedure we exclude those regional languages that are spoken by a very small percentage of the population of the country. The list of official languages is obtained from (Nations Online Project, 2016). We make an exception in Czech Republic where Slovenian is kind of an official language (Wikipedia, 2016) and we do not apply the 10 % rule in this case.

⁴³ Current prices, Gross domestic product at market prices

The distance between the countries is calculated using the library *cshapes* of R (Weidmann, Kuse, & Gleditsch, 2011). We include in the model the average distance and a dummy variable to indicate whether or not the countries share borders.

We include in the model the regulatory quality of the destination country using the Regulatory Quality variable in 2009 of the Worldwide Governance Indicator (World Bank, 2016). It is expected that people tend to move to more reliable countries.

We include the GDP per capita of the origin and destination countries in 2009 to control for the richness of the countries (European Commission, 2016g; EUROSTAT, 2016c).

We also include the unemployment rate of the origin and destination countries in 2009 (Eurostat, 2014). It is expected that people move from countries with higher unemployment rates to countries with lower unemployment rates where finding a job is expected to be easier.

7.5.1.4 Empirical model

Gravity models are usually estimated using a log-log model because the model becomes linear and the coefficient becomes elasticities indicating the percentage change of the response variable to the percentage change of the independent variable. In the case of a dummy or dummy-transformed factor variable the coefficient multiplied by 100 indicates the percentage change of the response variable when the dummy condition holds.

However, as described by Cardona et al. (2015) there is some controversy with these models because of the heteroskedasticity in the error term and some observations having zero values that are left out of the model when using log transformations. Therefore we use several models to provide more reliable results (Poisson, Poisson excluding the zero values, OLS excluding the zero values, OLS using fixed country dummies as suggested by (Feenstra, 2002), and a Tobit zero inflated model). We choose the results of the Poisson model that is likely to produce more accurate results in gravity models as demonstrated by (Silva & Tenreyro, 2006, 2011).

Our model in log-log form is as follows:

$$\begin{aligned} \log(p_{in\ o}) = & \beta_0 + \delta_{1:4} (l_{c\ o}^b) + \beta_1 \log(G_{a\ d}) + \\ & + \beta_2 \log(p_{o}) + \beta_3 \log(d_{o}) + \delta_5 (n_{hb\ ho\ o}) + \\ & \beta_4 \log(r_{q\ d}) + \beta_5 \log(G_{p\ c\ d}) + \beta_6 \log(G_{p\ c\ o}) + \\ & \beta_7 \log(u_{r\ d}) + \beta_8 \log(u_{r\ o}) + \varepsilon \end{aligned}$$

We estimate the standard errors of the results using robust standard errors. We also estimate the standard errors using bootstrapping with 10 000 interactions.

7.5.1.5 Results

The results of the analysis using robust standard errors and bootstrapping are shown in Table 21 and Table 22.

Having low language barriers increase the percentage of EU citizens that have lived and worked in a foreign EU country by 117.7 % and it is statistically significant. The variables related to the size of the countries are closed to 1 as expected. The distance has the expected sign while neighbourhood is not significant. The regulatory quality, GDP per capita of the origin country is significant and have the expected sign. Unemployment rate of the destination country is also significant but have an unexpected sign.

Table 21: Regression results for cross-border workers' mobility using Robust Standard Errors

Regression results for Cross-border Workers' Mobility						
	Dependent variable: Cross-border Workers					
	<i>Poisson</i>		<i>OLS</i>	<i>Poisson</i>	<i>Tobit</i>	
	y>0	y>0	y>0	Country dummy Coefficients	Adjusted	
	(1)	(2)	(3)	(4)	(5)	(6)
medium-high language barriers	0.630*** (0.237)	0.382 (0.247)	0.342 (0.211)	0.554*** (0.211)	0.745*** (0.224)	0.479
medium-low language barriers	0.658** (0.328)	0.567** (0.279)	0.392 (0.308)	0.391 (0.300)	0.671* (0.366)	0.432
low language barriers	1.183*** (0.279)	0.897*** (0.256)	0.908*** (0.298)	0.726*** (0.265)	1.344*** (0.308)	0.865
log(GDP destination country)	1.093*** (0.066)	0.778*** (0.069)	0.679*** (0.061)		1.011*** (0.073)	0.651
log(population)	0.892*** (0.071)	0.830*** (0.072)	0.771*** (0.052)		0.764*** (0.065)	0.491
log(distance)	-0.545*** (0.194)	-0.303 (0.201)	-0.228 (0.167)	-1.092*** (0.225)	-0.758*** (0.191)	-0.487
log(regulatory quality)	0.899** (0.442)	1.224*** (0.475)	1.529*** (0.399)		1.138** (0.493)	0.732
neighbour	0.312 (0.229)	0.399* (0.231)	0.570** (0.270)	0.157 (0.210)	0.257 (0.291)	0.165
GDP per capita origin country	-0.029*** (0.009)	-0.025*** (0.009)	-0.011** (0.006)		-0.015** (0.007)	-0.010
GDP per capita destination country	-0.003 (0.016)	-0.041*** (0.016)	-0.039*** (0.009)		-0.004 (0.011)	-0.002
unemployment rate origin country	0.033 (0.023)	0.028 (0.025)	-0.007 (0.020)		-0.004 (0.024)	-0.003
unemployment rate destination country	0.087*** (0.026)	0.047** (0.023)	0.039* (0.021)		0.092*** (0.026)	0.059
Constant	-5.258*** (0.713)	-3.288*** (0.697)	-3.464*** (0.535)	-1.163** (0.529)	-5.036*** (0.677)	-3.240
Observations	702	280	280	702	702	702
R ²			0.647			
Adjusted R ²			0.631			
Log Likelihood	-Inf.000	-Inf.000		-Inf.000	-374.332	-374.332
Akaike Inf. Crit.	Inf.000	Inf.000		Inf.000		
Residual Std. Error			1.090 (df = 267)			
F Statistic			40.830*** (df = 12; 267)			
Wald Test (df = 12)					420.833***	420.833***
<i>Note:</i>				* p<0.1; ** p<0.05; *** p<0.01		
				Robust se		
				Adjusted coefficients for Tobit (APE): 0.64		

Source: Compiled by the authors based on (European Commission, 2009, 2014a; EUROSTAT, 2016c)

Table 22: Regression results for cross-border workers' mobility using Bootstrapping (10 000 iterations)

Regression results for Cross-border Workers' Mobility						
	Dependent variable: Cross-border Workers					
	<i>Poisson</i>		<i>OLS</i>	<i>Poisson</i>	<i>Tobit</i>	
		y>0	y>0	Country dummy Coefficients	Adjusted	
	(1)	(2)	(3)	(4)	(5)	(6)
medium-high language barriers	0.630	0.382	0.342	0.554	0.745*	0.479
	p = 0.118	p = 0.231	p = 0.200	p = 0.196	p = 0.053	
medium-low language barriers	0.658	0.567	0.392	0.391	0.671	0.432
	p = 0.178	p = 0.170	p = 0.260	p = 0.334	p = 0.199	
low language barriers	1.183**	0.897**	0.908*	0.726	1.344**	0.865
	p = 0.028	p = 0.050	p = 0.072	p = 0.176	p = 0.030	
log(GDP destination country)	1.093***	0.778***	0.679***		1.011***	0.651
	p = 0.000	p = 0.000	p = 0.000		p = 0.000	
log(population)	0.892***	0.830***	0.771***		0.764***	0.491
	p = 0.000	p = 0.000	p = 0.000		p = 0.000	
log(distance)	-0.545*	-0.303	-0.228	-1.092**	-0.758**	-0.487
	p = 0.085	p = 0.225	p = 0.241	p = 0.022	p = 0.023	
log(regulatory quality)	0.899	1.224	1.529**		1.138	0.732
	p = 0.168	p = 0.106	p = 0.031		p = 0.126	
neighbour	0.312	0.399	0.570	0.157	0.257	0.165
	p = 0.284	p = 0.232	p = 0.159	p = 0.404	p = 0.342	
GDP per capita origin country	-0.029*	-0.025	-0.011		-0.015	-0.010
	p = 0.079	p = 0.110	p = 0.166		p = 0.146	
GDP per capita destination country	-0.003	-0.041	-0.039**		-0.004	-0.002
	p = 0.484	p = 0.105	p = 0.022		p = 0.449	
unemployment rate origin country	0.033	0.028	-0.007		-0.004	-0.003
	p = 0.283	p = 0.332	p = 0.434		p = 0.473	
unemployment rate destination country	0.087*	0.047	0.039		0.092**	0.059
	p = 0.055	p = 0.169	p = 0.173		p = 0.041	
Constant	-5.258***	-3.288**	-3.464***	-1.163	-5.036***	-3.240
	p = 0.0002	p = 0.012	p = 0.001	p = 0.176	p = 0.0001	
Observations	702	280	280	702	702	702
R ²			0.647			
Adjusted R ²			0.631			
Log Likelihood	-Inf.000	-Inf.000		-Inf.000	-374.332	-374.332
Akaike Inf. Crit.	Inf.000	Inf.000		Inf.000		
Residual Std. Error			1.090 (df = 267)			
F Statistic			40.830*** (df = 12; 267)			
Wald Test (df = 12)					420.833***	420.833***

Note:

*p<0.1; **p<0.05; ***p<0.01

Bootstrapping 10.000 interactions

Adjusted coefficients for Tobit (APE): 0.64

Source: Compiled by the authors based on (European Commission, 2009, 2014a; EUROSTAT, 2016c)

7.5.2 Analysis of the cross-border e-commerce (demand side)

We analyse the population making cross-border on-line purchases in 2011 among EU27 countries in 2011. Croatia is excluded of the analysis because there were no data available of this country in the Eurobarometer 77.1 (European Commission, 2014a). We create a dataset of 702 edges (27 origin countries x 27 destination countries excluding the pairs of a country to itself).

7.5.2.1 Response variable

The response variable is the number of on-line cross-border shoppers. To calculate this value we use the 2011⁴⁴ data of percentage of on-line shoppers buying in other country included in the study of Civic Consulting (2011). Based on this percentage and using data of population buying on-line from Eurostat (EUROSTAT, 2016a) we calculate the number of on-line cross-border shoppers among EU countries.

7.5.2.2 Independent variable

To calculate the language barriers between the countries we use the dataset of the Eurobarometer 77.1 that include a topic about Multilingualism (European Commission, 2014a). There are specific questions in the survey about the foreign languages spoken and the level of knowledge of the language (very good, good, and basic) in 2011. Based on that questions we classify the language similarities between the origin country and the destination country into four groups, namely low language barriers (more than 85 % of the population in the origin country speaks, at least good, one of the majority official languages of the destination country⁴⁵), medium-low language barriers (between 50 % and 85 %), medium-high language barriers (between 15 % and 50 %) and high language barriers (less than 15 %).

7.5.2.3 Control variables

The market size of the destination country is the GDP of the country in 2011⁴⁶ (EUROSTAT, 2016c). It is expected that GDP act as a proxy for the attractiveness of the e-market. Countries with bigger GDPs are likely to have a more competitive offering of e-shops.

We estimate the “market size” of the origin country using the population who buy on-line (EUROSTAT, 2016a). We think that these figures reflect more accurately than the GDP of the country the size of the on-line market of the origin country.

The distance between the countries is calculated using the library cshapes of R (Weidmann et al., 2011). We include in the model the average distance and a dummy variable to indicate whether or not the countries share borders.

We also include in the model the regulatory quality of the destination country using the regulatory quality variable in 2011 of the Worldwide Governance Indicator (World Bank, 2016). It is expected that people tend to buy in more reliable countries.

⁴⁴ We have seen that the cross-border e-commerce share in total e-commerce has not substantially changed since 2011 so the estimates are likely to be valid in 2016.

⁴⁵ We consider that the official national or regional language is a majority language in the country if it is spoken, with a level at least good, by more than 10 % of the population. Using this procedure we exclude those regional languages that are spoken by a very small percentage of the population of the country. The list of official languages is obtained from (Nations Online Project, 2016). We make an exception in Czech Republic where Slovenian is kind of an official language (Wikipedia, 2016) and we do not apply the 10 % rule in this case.

⁴⁶ Current prices, Gross domestic product at market prices

Eventually we include three dummy variables when the destination country is DE, UK, or FR because these three countries seem to be the favourite destination countries for cross-border online shopping (Civic Consulting, 2011).

7.5.2.4 Empirical model

Gravity models are usually estimated using a log-log model because the model becomes linear and the coefficient becomes elasticities indicating the percentage change of the response variable to the percentage change of the independent variable. In the case of a dummy or dummy-transformed factor variable the coefficient multiplied by 100 indicates the percentage change of the response variable when the dummy condition holds.

However, as described by Cardona et al. (2015) there is some controversy with these models because of the heteroskedasticity in the error term and some observations having zero values that are left out of the model when using log transformations. Therefore we use several models to provide more reliable results (Poisson, Poisson excluding the zero values, OLS excluding the zero values, OLS using fixed country dummies as suggested by (Feenstra, 2002), and a Tobit zero inflated model). We choose the results of the Poisson model that is likely to produce more accurate results in gravity models as demonstrated by (Silva & Tenreyro, 2006, 2011).

Our model in log-log form is as follows:

$$\log(p_{sho_{o,a}}) = \beta_0 + \delta_{1:4}(\ln \frac{b_o}{a_o}) + \beta_1 \log(G_d) + \beta_2 \log(p_{sho_{o,o}}) + \beta_3 \log(d_o) + \delta_5(n_{hb_{ho,o}}) + \beta_4 \log(r_{q,a}) + \delta_6(D_d) + \delta_7(U_d) + \delta_8(F_d) + \varepsilon$$

We estimate the standard errors of the results using robust standard errors. We also estimate the standard errors using bootstrapping with 10 000 interactions.

7.5.2.5 Results

The results of the analysis using Robust Standard Errors and bootstrapping are shown in Table 23 and Table 24.

Having low language barriers increases 142 % the number of on-line shoppers buying cross-border compared to having high language barriers. That means that the number of cross-border shoppers is 4.14 ($\exp(1.42)$) times higher between countries with low language barriers compared to high language barriers. The coefficient of the market size of the countries is close to 1 as expected. Regarding the distance and whether or not two countries are neighbours we see that the effect of distance is very small while the effect of sharing the borders is more relevant (44 % increase in the number of on-line cross-border shoppers between neighbouring countries). The effect is still much lower than the barrier language, suggesting that the buyers are not concerned with the distance of the destination country but with being able to effectively communicate with the foreign shop. The coefficient of the regulatory variable is also significant. It is likely that people prefer to shop in more reliable countries. The dummy countries for DE, UK, and FR are acting as expected, although only DE and UK are significant.

Table 23: Regression results for cross-border on-line shoppers

Regression results for cross-border on-line shoppers						
	Dependent variable: Cross-border online-shoppers					
	<i>Poisson</i>		<i>OLS</i>	<i>Poisson</i>	<i>Tobit</i>	
	(1)	y>0 (2)	y>0 (3)	Country dummy (4)	Coefficients (5)	Adjusted (6)
medium-high language barriers	0.160 (0.180)	0.057 (0.164)	0.236** (0.111)	0.349** (0.168)	0.796*** (0.181)	0.611
medium-low language barriers	0.341 (0.221)	0.267 (0.208)	0.406** (0.165)	0.679*** (0.200)	1.050*** (0.179)	0.806
low language barriers	1.420*** (0.300)	1.232*** (0.306)	1.216*** (0.257)	1.375*** (0.223)	2.013*** (0.419)	1.544
log(GDP destination country)	0.810*** (0.095)	0.457*** (0.106)	0.338*** (0.044)		1.331*** (0.060)	1.021
log(population shopping on-line)	1.021*** (0.045)	0.990*** (0.044)	1.022*** (0.019)		1.005*** (0.037)	0.771
log(distance)	-0.259* (0.135)	-0.246* (0.130)	-0.326*** (0.071)	-0.675*** (0.145)	-0.364*** (0.121)	-0.279
log(regulatory quality)	0.977*** (0.149)	0.423*** (0.141)	0.139* (0.084)		1.216*** (0.198)	0.933
neighbour	0.439*** (0.156)	0.398*** (0.152)	0.562*** (0.138)	0.317** (0.155)	0.641*** (0.224)	0.492
DE	0.371* (0.193)	1.026*** (0.237)	1.546*** (0.141)		-0.242 (0.179)	-0.186
UK	0.620** (0.274)	1.223*** (0.286)	1.937*** (0.174)		0.080 (0.221)	0.062
FR	0.201 (0.268)	0.581** (0.278)	0.791*** (0.139)		-0.317* (0.164)	-0.243
Constant	-3.349*** (0.347)	-3.124*** (0.339)	-3.713*** (0.134)	1.787*** (0.260)	-3.464*** (0.237)	-2.657
Observations	702	366	366	702	702	702
R ²			0.920			
Adjusted R ²			0.917			
Log Likelihood	-Inf.000	-Inf.000		-Inf.000	-607.874	-607.874
Akaike Inf. Crit.	Inf.000	Inf.000		Inf.000		
Residual Std. Error			0.592 (df = 354)			
F Statistic			368.336*** (df = 11; 354)			
Wald Test (df = 11)					1,832.426***	1,832.426***
<i>Note:</i>					*p<0.1; **p<0.05; ***p<0.01	
					Robust se	
					Adjusted coefficients for Tobit (APE): 0.77	

Source: Compiled by the authors based on (European Commission, 2009, 2014a; EUROSTAT, 2016c)

Table 24: Regression results for cross-border on-line shoppers using Bootstrapping (10 000 iterations)

Regression results for cross-border on-line shoppers						
	Dependent variable: Cross-border online-shoppers					
	<i>Poisson</i>		<i>OLS</i>	<i>Poisson</i>	<i>Tobit</i>	
	(1)	(2)	(3)	Country dummy	Coefficients	Adjusted
		y>0	y>0			
medium-high language barriers	0.160	0.057	0.236	0.349	0.796**	0.611
	p = 0.360	p = 0.460	p = 0.147	p = 0.220	p = 0.023	
medium-low language barriers	0.341	0.267	0.406	0.679	1.050***	0.806
	p = 0.250	p = 0.276	p = 0.110	p = 0.144	p = 0.004	
low language barriers	1.420**	1.232**	1.216**	1.375*	2.013***	1.544
	p = 0.016	p = 0.031	p = 0.011	p = 0.060	p = 0.010	
log(GDP destination country)	0.810***	0.457**	0.338***		1.331***	1.021
	p = 0.000	p = 0.016	p = 0.0001		p = 0.000	
log(population shopping on-line)	1.021***	0.990***	1.022***		1.005***	0.771
	p = 0.000	p = 0.000	p = 0.000		p = 0.000	
log(distance)	-0.259	-0.246	-0.326**	-0.675**	-0.364*	-0.279
	p = 0.166	p = 0.173	p = 0.011	p = 0.020	p = 0.075	
log(regulatory quality)	0.977***	0.423*	0.139		1.216***	0.933
	p = 0.0003	p = 0.080	p = 0.199		p = 0.0004	
neighbour	0.439	0.398	0.562**	0.317	0.641*	0.492
	p = 0.105	p = 0.118	p = 0.024	p = 0.201	p = 0.082	
DE	0.371	1.026**	1.546***		-0.242	-0.186
	p = 0.182	p = 0.018	p = 0.000		p = 0.267	
UK	0.620	1.223**	1.937***		0.080	0.062
	p = 0.131	p = 0.018	p = 0.000		p = 0.429	
FR	0.201	0.581	0.791***		-0.317	-0.243
	p = 0.349	p = 0.149	p = 0.003		p = 0.176	
Constant	-3.349***	-3.124***	-3.713***	1.787**	-3.464***	-2.657
	p = 0.000	p = 0.000	p = 0.000	p = 0.016	p = 0.000	
Observations	702	366	366	702	702	702
R ²			0.920			
Adjusted R ²			0.917			
Log Likelihood	-Inf.000	-Inf.000		-Inf.000	-607.874	-607.874
Akaike Inf. Crit.	Inf.000	Inf.000		Inf.000		
Residual Std. Error			0.592 (df = 354)			
F Statistic			368.336*** (df = 11; 354)			
Wald Test (df = 11)					1,832.426***	1,832.426***
<i>Note:</i>				* p<0.1; ** p<0.05; *** p<0.01		
				Bootstrapping 10.000 interactions		
				Adjusted coefficients for Tobit (APE): 0.77		

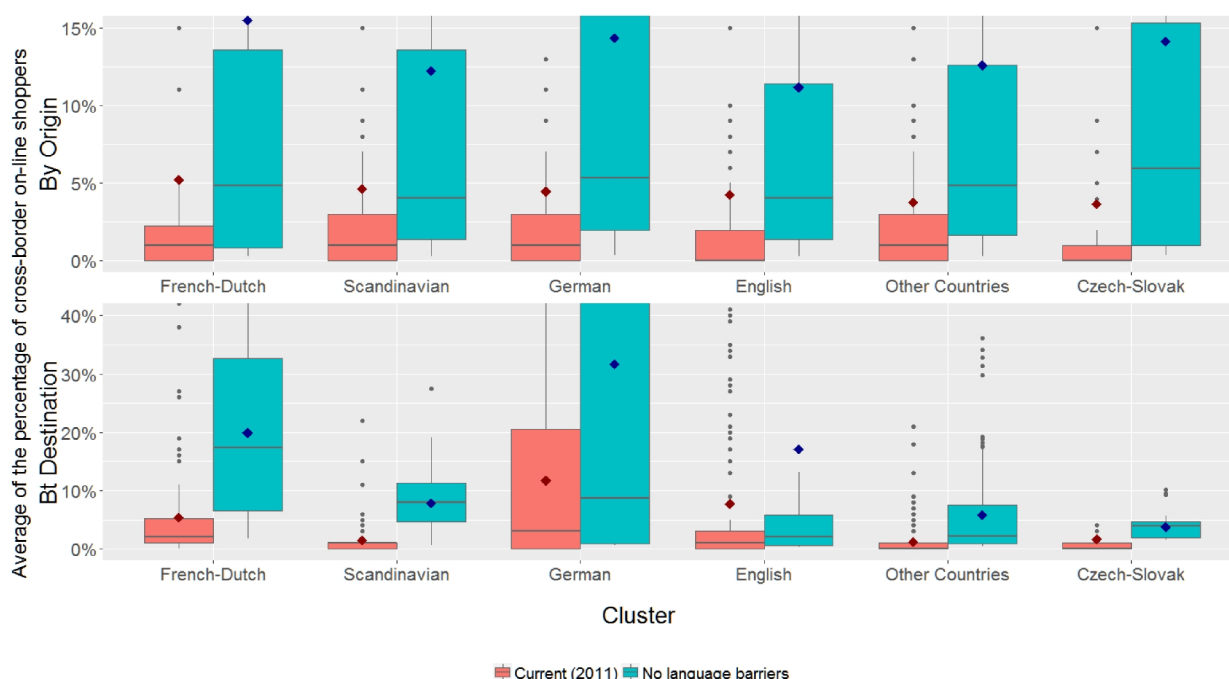
Source: Compiled by the authors based on (European Commission, 2009, 2014a; EUROSTAT, 2016c)

7.5.2.6 Simulation of the effect of not having language barriers on cross-border on-line shoppers

Using these results we can estimate what would happen in a theoretical scenario if high quality HLT could allow a simple and efficient mean to automatic translation of the e-commerce websites. In this scenario we could consider that the language barriers between all the countries were low, because all the web-sites would be automatically translated to the preferred languages of the consumers, the search engines could easily refer to those pages, and the post-sale service could be also provided in the language of the consumer.

The average percentage of shoppers who buy on line abroad out of the total number of on-line shoppers (by origin and destination country) will climb from the current 4.3 % to 13 %. More interesting is to make the analysis by using the groups of the current fragmented market. We have estimated the average percentage of European e-commerce consumers buying in another EU country by group in the current situation and in a theoretical situation where the language barriers would be overcome. The results shown in Figure 53 suggest that the percentage of users will grow in a much more balanced way. The average percentage of users buying to other countries would substantially increase in all the groups (up to between 11 % and 15 %) while in the current situation, all the values are below 5 % with the countries in the “Other countries” group clearly left behind (3.8 %). Overcoming barrier languages will foster a much higher and efficient market that will particularly benefit countries that do not share the language with other European countries.

Figure 53: Percentage of cross-border online-shoppers of total on-line shoppers by cluster and depending on language barriers (box-plot and mean)



Source: Compiled by the authors based on (Civic Consulting, 2011; European Commission, 2014a; EUROSTAT, 2016c)

The effect is also relevant when analysing the destination countries of the e-shoppers. Big economies that are likely to have more attractive and competitive on-line retail markets will be the most benefited. However, even countries that currently receive negligible cross-border e-commerce traffic would see a

substantial increase of average percentage of users coming from other countries (from 1 % in the current situation up to 6 % for the countries in the “Other Countries” group).

7.5.3 Analysis of the cross-border e-commerce (supply side)

We analyse the retailers selling on-line cross-border in 2015 among EU27 countries in 2011. Croatia is excluded of the analysis because there were no data available of this country in the Eurobarometer 77.1 (European Commission, 2014a). We create a dataset of 702 edges (27 origin countries x 27 destination countries excluding the pairs of a country to itself).

7.5.3.1 Response variable

The response variable is the number of on-line cross-border retailers selling to individual consumers in 2015. To calculate this value we use the dataset of the Flash Eurobarometer 413 – Companies Engaged in Online Activities – (European Commission, 2015c) that includes information about the countries where the retailer is selling.

7.5.3.2 Independent variable

To calculate the language barriers between the countries we use the dataset of the Eurobarometer 77.1 that include a topic about Multilingualism (European Commission, 2014a). There are specific questions in the survey about the foreign languages spoken and the level of knowledge of the language (very good, good, and basic) in 2011. Based on that questions we classify the language similarities between the origin country and the destination country into four groups, namely low language barriers (more than 85 % of the population in the origin country speaks, at least good, one of the majority official languages of the destination country⁴⁷), medium-low language barriers (between 50 % and 85 %), medium-high language barriers (between 15 % and 50 %) and high language barriers (less than 15 %). We are using data of 2011 because this is the only available dataset that provides rich information about foreign languages. Although using data of 2011 when analysing the effect on a variable of 2015 may introduce some error we expect that the results are still reliable because we have seen that the changes in the percentage of speakers of a foreign language in a country are very likely to evolve very slowly (see Figure 15).

7.5.3.3 Control variables

The market size of the destination country is the GDP of the country in 2011⁴⁸ (EUROSTAT, 2016c). It is expected that GDP act as a proxy for the attractiveness of the e-market. Countries with bigger GDPs are likely to have a more competitive offering of e-shops.

We estimate the “market size” of the origin country using the retailers who sell on-line to individual consumers of the dataset of the Flash Eurobarometer 413 (Companies Engaged in Online Activities) (European Commission, 2015c). We think that these figures reflect more accurately than the GDP of the country the size of the on-line retailer market of the origin country.

⁴⁷ We consider that the official national or regional language is a majority language in the country if it is spoken, with a level at least good, by more than 10 % of the population. Using this procedure we exclude those regional languages that are spoken by a very small percentage of the population of the country. The list of official languages is obtained from (Nations Online Project, 2016). We make an exception in Czech Republic where Slovenian is kind of an official language (Wikipedia, 2016) and we do not apply the 10 % rule in this case.

⁴⁸ Current prices, Gross domestic product at market prices

The distance between the countries is calculated using the library *cshapes* of R (Weidmann et al., 2011). We include in the model the average distance and a dummy variable to indicate whether or not the countries share borders.

We also include in the model the regulatory quality of the destination country using the regulatory quality variable in 2015 of the Worldwide Governance Indicator (World Bank, 2016). It is expected that people tend to buy in more reliable countries.

In this case we do not include the dummy countries for DE, UK, and FR because we find no evidence that there is a strong preference for selling to these markets.

7.5.3.4 Empirical model

Gravity models are usually estimated using a log-log model because the model becomes linear and the coefficient becomes elasticities indicating the percentage change of the response variable to the percentage change of the independent variable. In the case of a dummy or dummy-transformed factor variable the coefficient multiplied by 100 indicates the percentage change of the response variable when the dummy condition holds.

However, as described by Cardona et al. (2015) there is some controversy with these models because of the heteroskedasticity in the error term and some observations having zero values that are left out of the model when using log transformations. Therefore we use several models to provide more reliable results (Poisson, Poisson excluding the zero values, OLS excluding the zero values, OLS using fixed country dummies as suggested by (Feenstra, 2002), and a Tobit zero inflated model). We choose the results of the Poisson model that is likely to produce more accurate results in gravity models as demonstrated by (Silva & Tenreyro, 2006, 2011).

Our model in log-log form is as follows:

$$\log(p_{sho_{o,a}}) = \beta_0 + \delta_{1:4} (l_{b_o}) + \beta_1 \log(G_a) + \beta_2 \log(p_{sho_{o,o}}) + \beta_3 \log(d_{o_o}) + \delta_5 (n_{hb_{ho_o}}) + \beta_4 \log(r_{q_a,a}) + \varepsilon$$

We estimate the standard errors of the results using robust standard errors. We also estimate the standard errors using bootstrapping with 10.000 interactions.

7.5.3.5 Results

The results of the analysis using Robust Standard Errors and bootstrapping are shown in Table 25 and Table 26.

Logistics seem to have a significant effect because the neighbour factor is higher than in the consumers' model (60% increase compared to 44%) and the distance effect is much higher and very significant. This is in line with the barriers described in Figure 27. Still, the percentage of shops that sells on-line only is higher when the language barriers are very low (1.6 times higher, $\exp(0.468)$). The regulatory quality of the destination country becomes irrelevant. All in all, it seems that companies consider selling to another country when the destination country is close to the country of the seller and shares the same language.

Table 25: Regression results for cross-border on-line retailers

Regression results for cross-border on-line retailers						
	Dependent variable: Cross-border online-shops selling to individuals					
	<i>Poisson</i>		<i>OLS</i>	<i>Poisson</i>	<i>Tobit</i>	
	y>0	y>0	y>0	Country dummy	Coefficients	Adjusted
	(1)	(2)	(3)	(4)	(5)	(6)
medium-high language barriers	-0.139 (0.194)	-0.110 (0.186)	-0.092 (0.223)	0.552** (0.238)	-0.110 (0.256)	-0.079
medium-low language barriers	-0.198 (0.174)	-0.026 (0.164)	0.210 (0.310)	0.339* (0.181)	-0.164 (0.360)	-0.117
low language barriers	0.468*** (0.148)	0.574*** (0.141)	0.731*** (0.197)	0.706*** (0.220)	0.564** (0.267)	0.402
log(GDP destination country)	0.396*** (0.085)	0.352*** (0.082)	0.350*** (0.072)		0.503*** (0.081)	0.359
log(companies selling on-line)	1.078*** (0.104)	1.059*** (0.101)	1.037*** (0.078)		1.016*** (0.096)	0.725
log(regulatory quality)	0.101 (0.159)	0.088 (0.152)	0.192 (0.142)		0.059 (0.178)	0.042
log(distance)	-0.608*** (0.120)	-0.401*** (0.124)	-0.235* (0.128)	-0.856*** (0.100)	-0.595*** (0.140)	-0.424
neighbour	0.599*** (0.161)	0.665*** (0.152)	1.237*** (0.193)	0.104 (0.124)	1.003*** (0.188)	0.715
Constant	-3.042*** (0.463)	-2.887*** (0.449)	-3.364*** (0.346)	-0.828*** (0.267)	-3.270*** (0.431)	-2.333
Observations	650	519	519	650	650	650
R ²			0.715			
Adjusted R ²			0.710			
Log Likelihood	-Inf.000	-Inf.000		-Inf.000	-395.461	-395.461
Akaike Inf. Crit.	Inf.000	Inf.000		Inf.000		
Residual Std. Error			1.197 (df = 510)			
F Statistic			159.777*** (df = 8; 510)			
Wald Test (df = 8)					749.561***	749.561***

Note:

*p<0.1; **p<0.05; ***p<0.01

Robust se

Adjusted coefficients for Tobit (APE): 0.71

Source: Compiled by the authors based on (European Commission, 2009, 2014a; EUROSTAT, 2016c)

Table 26: Regression results for cross-border on-line retailers using Bootstrapping (10 000 iterations)

Regression results for cross-border on-line retailers						
	Dependent variable: Cross-border online-shops selling to individuals					
	<i>Poisson</i>		<i>OLS</i>	<i>Poisson</i>	<i>Tobit</i>	
	y>0	y>0	y>0	Country dummy	Coefficients	Adjusted
	(1)	(2)	(3)	(4)	(5)	(6)
medium-high language barriers	-0.139	-0.110	-0.092	0.552	-0.110	-0.079
	p = 0.355	p = 0.374	p = 0.423	p = 0.209	p = 0.419	
medium-low language barriers	-0.198	-0.026	0.210	0.339	-0.164	-0.117
	p = 0.297	p = 0.480	p = 0.342	p = 0.317	p = 0.418	
low language barriers	0.468*	0.574**	0.731**	0.706	0.564	0.402
	p = 0.078	p = 0.033	p = 0.043	p = 0.166	p = 0.160	
log(GDP destination country)	0.396***	0.352**	0.350***		0.503***	0.359
	p = 0.008	p = 0.013	p = 0.007		p = 0.001	
log(companies selling on-line)	1.078***	1.059***	1.037***		1.016***	0.725
	p = 0.000	p = 0.000	p = 0.000		p = 0.000	
log(regulatory quality)	0.101	0.088	0.192		0.059	0.042
	p = 0.369	p = 0.378	p = 0.253		p = 0.426	
log(distance)	-0.608***	-0.401*	-0.235	-0.856***	-0.595**	-0.424
	p = 0.006	p = 0.058	p = 0.184	p = 0.0004	p = 0.018	
neighbour	0.599**	0.665**	1.237***	0.104	1.003***	0.715
	p = 0.041	p = 0.019	p = 0.001	p = 0.303	p = 0.006	
Constant	-3.042***	-2.887***	-3.364***	-0.828*	-3.270***	-2.333
	p = 0.001	p = 0.001	p = 0.000	p = 0.051	p = 0.000	
Observations	650	519	519	650	650	650
R ²			0.715			
Adjusted R ²			0.710			
Log Likelihood	-Inf.000	-Inf.000		-Inf.000	-395.461	-395.461
Akaike Inf. Crit.	Inf.000	Inf.000		Inf.000		
Residual Std. Error			1.197 (df = 510)			
F Statistic			159.777*** (df = 8; 510)			
Wald Test (df = 8)					749.561***	749.561***

Note:

*p<0.1; **p<0.05; ***p<0.01

Bootstrapping 10.000 interactions

Adjusted coefficients for Tobit (APE): 0.71

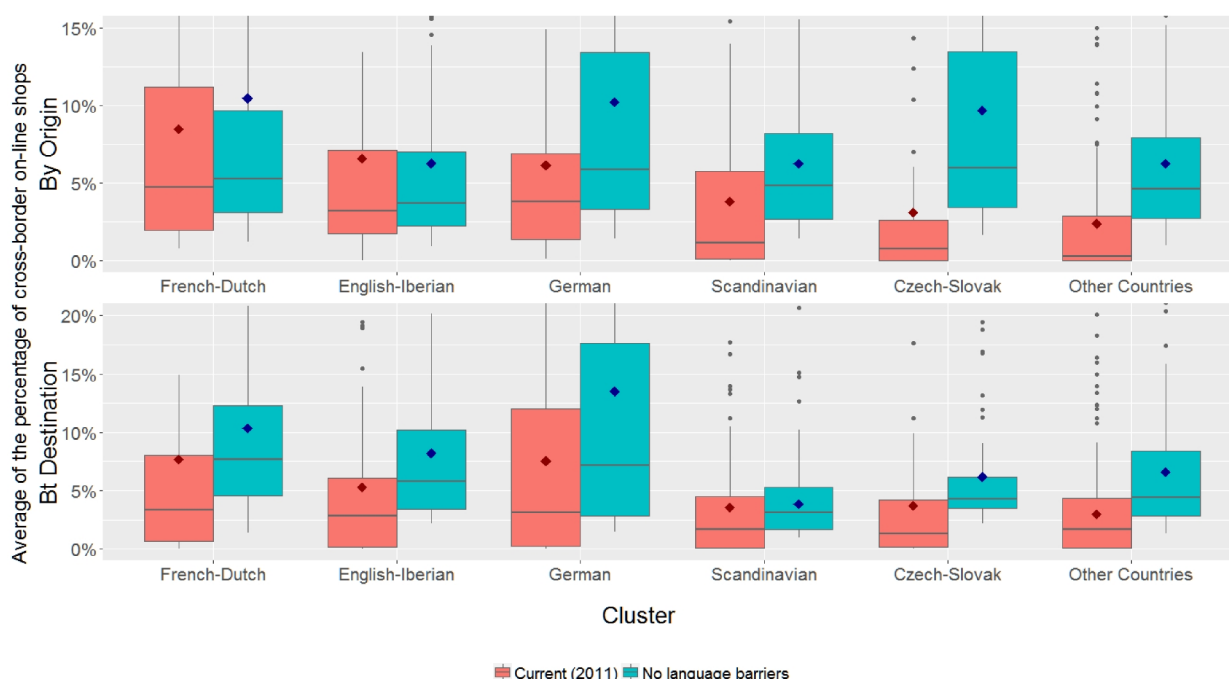
Source: Compiled by the authors based on (European Commission, 2009, 2014a; EUROSTAT, 2016c)

7.5.3.6 Simulation of the effect of not having language barriers on cross-border on-line sellers

Using these results we can estimate what would happen in a theoretical scenario if high quality HLT could allow a simple and efficient mean to automatic translation of the e-commerce websites. In this scenario we could consider that the language barriers between all the countries were low, because all the web-sites would be automatically translated to the preferred languages of the consumers, the search engines could easily refer to those pages, and the post-sale service could be also provided in the language of the consumer.

The average percentage of sellers who sell on line abroad, out of the total number of web sellers (by origin and destination country), will increase from the current 5 % to an estimated 7.6 %. The results by origin and destination countries are shown in Figure 54. Again, the average percentage of companies selling on-line to another EU country will increase in all the groups. Interestingly, the increase will particularly benefit those countries in the groups that currently have the lower percentage of e-retailers selling abroad. The average percentage of companies selling on-line to other countries would be higher than 6 % in all the groups while in the current situation, countries in the “Other countries” group are clearly left behind (2 % compared to 8 % in the French-Dutch group).

Figure 54: Country average of the percentage of cross-border online-shops of total on-line shops by cluster and depending on language barriers (box-plot and mean)



Source: Compiled by the authors based on (European Commission, 2009, 2014a; EUROSTAT, 2016c)

7.5.4 Limitations

The main limitation is to consider only one year when making the analysis (2009 for workers' mobility, 2011 for e-shoppers, and 2015 for web retailers). Instead of capturing the average effect, the results might be related to specificities of the year that could bias the estimate of the coefficients. However, there are no longitudinal data-sets to allow making a reliable multi-year analysis.

Another limitation is to consider the language barriers in 2011 when analysing workers' mobility in 2009 and web retailers in 2015. Although it may introduce some bias, we expect that the results are still reliable because the changes in the percentage of speakers of a foreign language in a country are very likely to evolve slowly.

The results should be taken carefully because language similarities may be related not only to the language itself but to other cultural similarities not captured in the model, or higher levels of trust between the countries for historical reasons.

Eventually, considering the demand and supply models separately is likely to be inadequate. Supply and demand are simultaneous in nature and usually they affect each other. This can be solved by using

simultaneous equations and methods such as two stage least squares with instrumental variables to solve the equations. In this specific case, defining and solving such model is problematic, because the observations are from different years and because the model is solved by using a Poisson regression.

7.6 Text-mining analysis

The documents analysed are the following:

- Technical documents of the DG Connect from the inventory of the reports on the studies completed by the European Commission Directorate General for Communications Networks, Content and Technology (478 documents from 2003 to 2015) (DG Connect, 2014)
- Blog of the DSM of the DG Connect (440 posts from 06/04/2011 to 21/09/2016) (European Commission, 2015e)
- Blogs of the EP (2.006 posts from 02/03/2012 to 28/09/2016) (European Parliament, 2015)

The terms related to language technologies that have been included in the text-mining analysis are:

diversity language
diversity linguist
minority language
multilingual
language need
language challenge
language barrier
automat translat
speech recogn
natural language
language AND technology
language AND data
language AND big data
language AND machine learning
human language technology

And the terms related to other trending technology topics are:

e government
cloud computing
smartphone
wearable
internet of the things
smart cities
big data
machine learning
open data

It may be questioned whether the terms have been chosen correctly. Assuming the important limitations of the analysis, we still think that they give an overall good idea about the differences between multilingualism and HLT topics compared to other trending technology topics.

A loess smoothing method has been used to plot the trends by using the function `stat_smooth` (default values) of the `ggplot` package (Wickham, 2009).

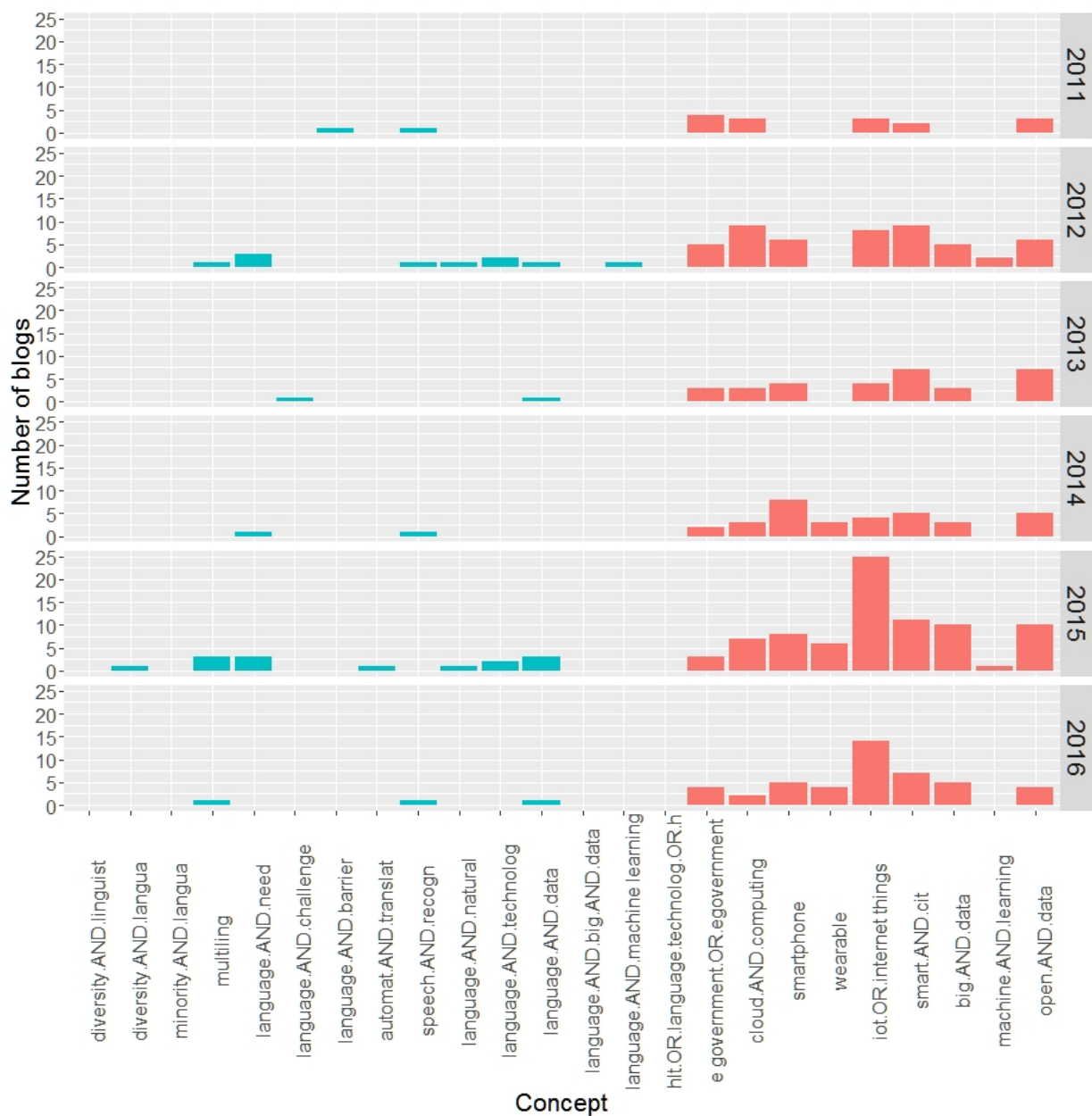
7.6.1 Number of documents by term and evolution

The following figures show the number of documents including each of the terms by year for the three groups.

Figure 55: Number of reports of the DG CONNECT by term and year

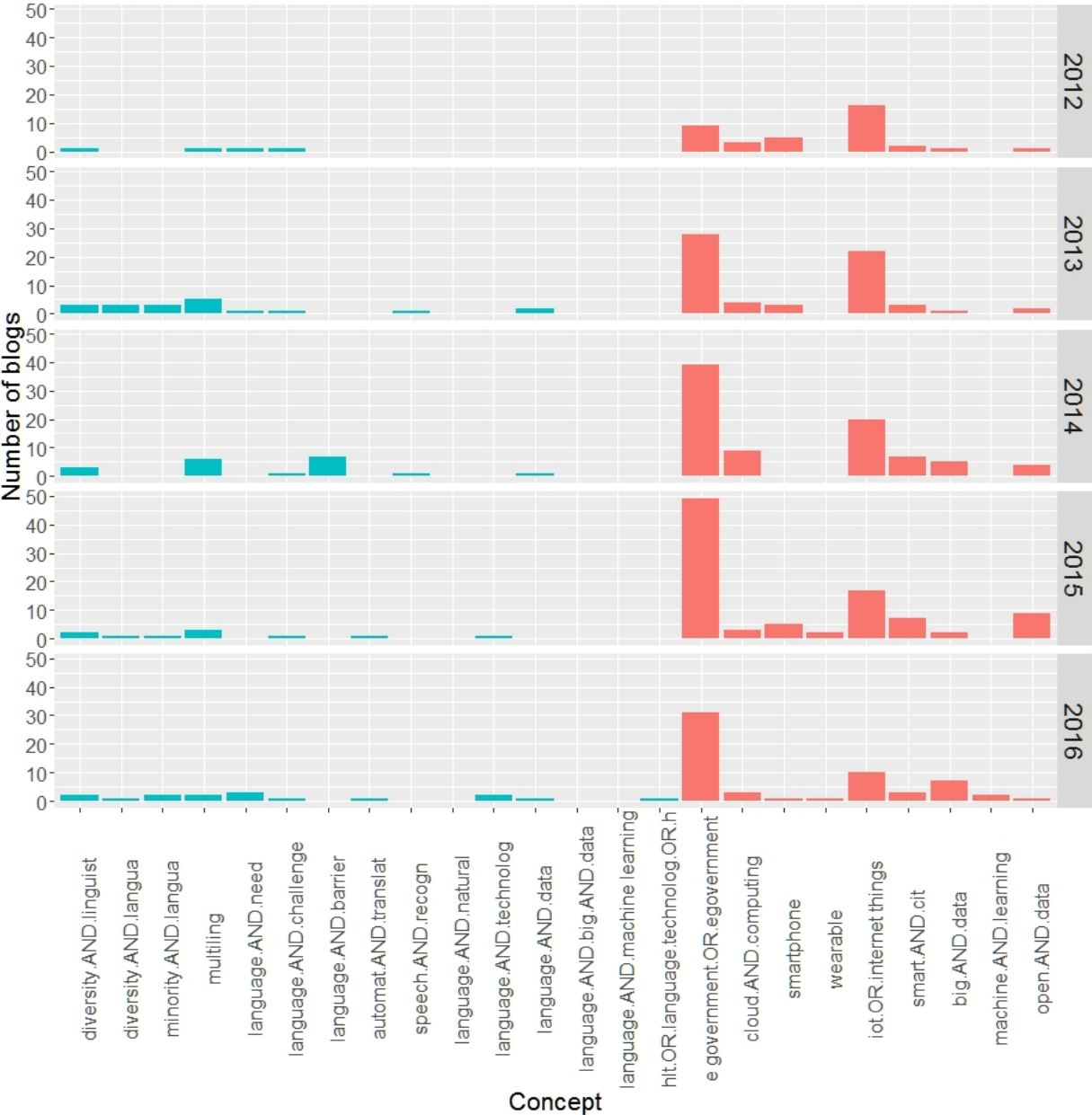


Source: Compiled by the authors based on (DG CONNECT, 2014)

Figure 56: Number of blogs of the Digital Single Market of the DG Connect by concept and year

Source: Compiled by the authors based on (European Commission, 2015e)

Figure 57: Number of blogs of the European Parliamentary Research Service by concept and year

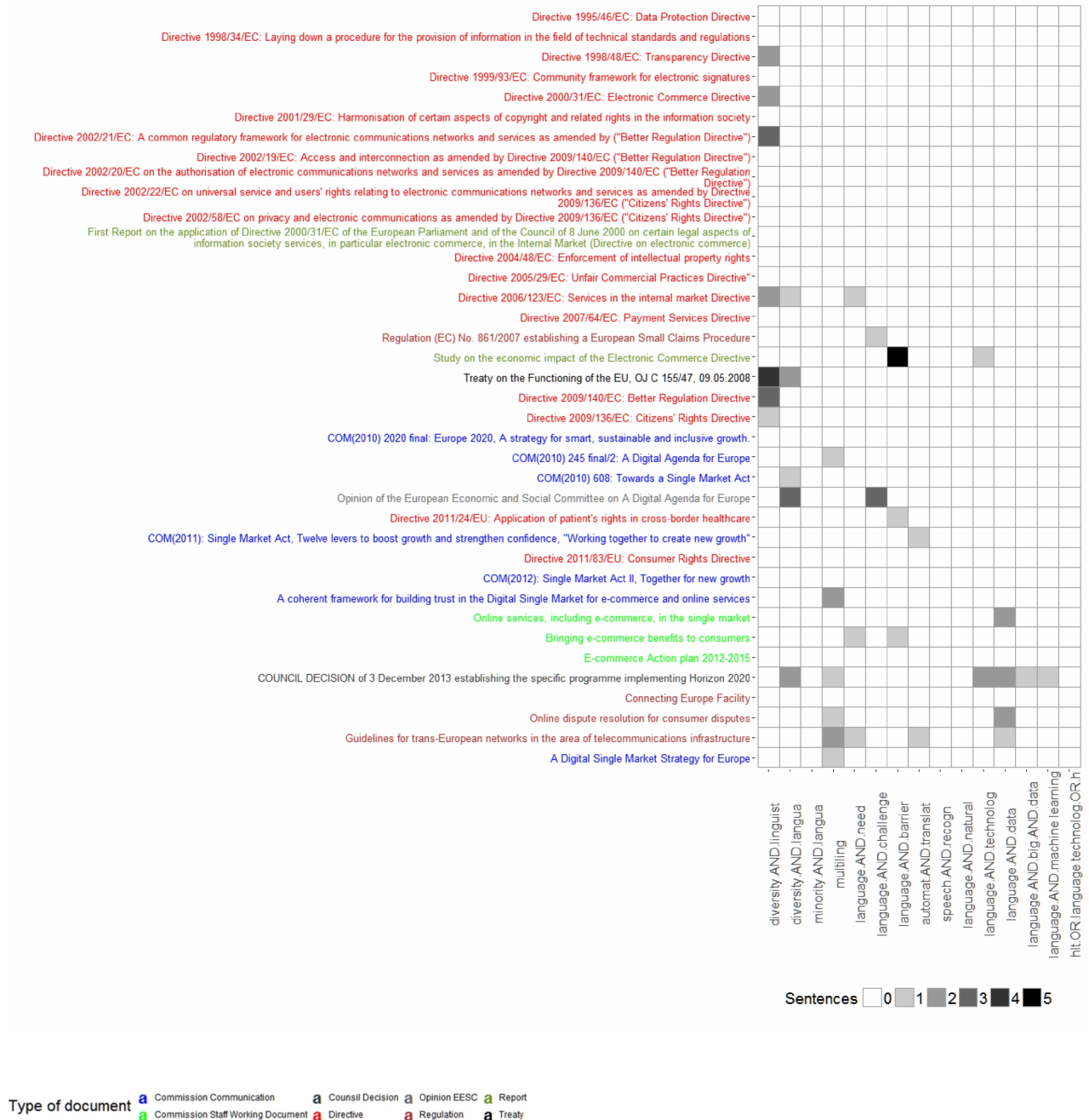


Source: Compiled by the authors based on (European Parliament, 2015)

7.6.2 List of official EU documents analysed

The following figure shows the list of official EU documents analysed and the number of sentences including the different HLT topics by document.

Figure 58: List of Official EU documents analysed and number of sentences including the topic



Source: Compiled by the authors

7.7 EU contribution to HLT related projects in the H2020, ICT-PSP and FP7 programs

The EU contribution to HLT related projects in the 7th FP, H2020, and ICT-PSP has been estimated using data provided by the EC and the dataset of projects funded by the EU under the 7th FP (European Commission, 2015h) for research and technological development. The list of projects and the corresponding EC funding is as follows:

Table 27: HLT related projects FP7

Acronym	Title	Maximum EC contribution
ACCEPT	Automated Community Content Editing PorTal	1.825.000 €
ACCURAT	Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation	2.825.000 €
AnnoMarket	Annotation Resource Marketplace in the Cloud	1.430.000 €
CASMACAT	Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation	2.500.000 €
CLARIN	Common Language Resources and Technology Infrastructure	4.100.000 €
CLASSIC	Computational Learning in Adaptive Systems for Spoken Conversation	3.400.000 €
CoSyne	Multi-Lingual Content Synchronization for Wikis	2.350.000 €
Dicta-Sign	Sign Language Recognition, Generation and Modelling \n with application in Deaf Communication	3.010.879 €
DIRHA	Distant-speech Interaction for Robust Home Applications	3.450.000 €
DISCOTEX	Distributional Compositional Semantics for Text Processing	1.087.930 €
EMIME	Effective Multilingual Interaction in Mobile Environments	3.050.000 €
EU-BRIDGE	Bridges Across the Language Divide	7.875.000 €
EUCases	EUropean and National CASE Law and Legislation Linked in Open Data Stack	1.499.000 €
EUMSSI	EUMSSI- Event Understanding through Multimodal Social Stream Interpretation	2.480.000 €
EuroMatrixPlus	Bringing Machine Translation for European Languages to the User	4.266.896 €
EUROSENTIMENT	Language Resource Pool for Sentiment Analysis in European Languages	1.930.000 €
EXCITEMENT	EXploring Customer Interactions through Textual EntailMENT	3.500.000 €
EXPERT	EXploiting Empirical appRoaches to Translation	3.935.340 €
FAUST	Feedback Analysis for User adaptive Statistical Translation	2.850.000 €
FIRST	A Flexible Interactive Reading Support Tool	2.008.754 €
GET HOME SAFE	Get Home Safe: Extended Multimodal Search and Communication Systems for Safe In-Car Application	3.100.000 €
LANGTERRA	Enhancing the Research Potential of ILSIP/"Athena" R.C. in Language Technology in the European Research ERA	1.689.320 €
LATEST	Advanced LAnguage TEchnology Platform for TranSLaTors (LATEST)	223.002 €
LIDER	LIDER: : Linked Data as an enabler of cross-media and multilingual content analytics for enterprises across Europe	1.482.000 €
LiMoSiNe	Linguistically Motivated Semantic aggregation engiNes	2.500.000 €
LT-Web	Language Technology in the Web	2.550.000 €
LT COMPASS	Guiding Language Technology Paths from Research to Markets	2.950.000 €
LTfLL	Language Technology for Lifelong Learning	2.849.604 €
MANTRA	Multilingual Annotation of Named Entities and Terminology Resources Acquisition	1.799.888 €
MateCat	Machine Translation Enhanced Computer Assisted Translation	2.650.000 €
MEDAR	Mediterranean Arabic Language and Speech Technology	798.552 €
METALOGUE	Multiperspective Multimodal Dialogue: dialogue system with metacognitive abilities	2.971.000 €
MICO	Media in Context	3.452.000 €
MLI	Towards a MultiLingual Data Services infrastructure	2.098.000 €
MOLTO	Multilingual On-Line Translation	2.975.000 €
MONNET	Multilingual Ontologies for Networked Knowledge	2.362.622 €
MosesCore	Moses Open Source Evaluation and Support Co-ordination for OutReach and Exploitation	1.200.000 €
MULTISENSOR	Mining and Understanding of multilingual contentT for Intelligent Sentiment Enriched coNtext and Social Oriented inteRpretation	2.965.000 €
OpenER	Open Polarity Enhanced Named Entity Recognition	1.930.000 €
ORGANIC	Self-organized recurrent neural learning for language processing	2.700.000 €
PANACEA	Platform for Automatic, Normalized Annotation and Cost-Effective Acquisition of Language Resources for Human Language Technologies	2.685.000 €
PARLANCE	Probabilistic Adaptive Real-Time Learning And\nNatural Conversational Engine	3.625.000 €
PHEME	Computing Veracity Across Media, Languages, and Social Networks	2.916.000 €
PortDial	Language Resources for Portable Multilingual Spoken Dialogue Systems	1.874.040 €
PRESEMT	Pattern Recognition-based Statistically Enhanced MT	2.500.000 €
QTLaunchPad	PREPARATION AND LAUNCH OF A LARGE-SCALE ACTION FOR QUALITY TRANSLATION TECHNOLOGY	1.914.000 €
QTLeap	Quality Translation by Deep Language Engineering Approaches	3.003.000 €
ROCKIT	Roadmap for Conversational Interaction Technologies	1.488.000 €
SAVAS	Sharing AudioVisual language resources for Automatic Subtitling	1.978.000 €
SENSEI	Making Sense of Human-Human Conversation Data	2.650.000 €
SIGNSPEAK	SCIENTIFIC UNDERSTANDING AND VISION-BASED TECHNOLOGICAL DEVELOPMENT FOR CONTINUOUS SIGN LANGUAGE RECOGNITION AND TRANSLATION	2.756.905 €
SIMPLE4ALL	Speech synthesis that improves through adaptive learning	3.100.000 €
smeSpire	A European Community of SMEs built on Environmental Digital Content and Languages	1.791.000 €
T4ME Net	Technologies for the Multilingual European Information Society	5.990.000 €
TaaS	Terminology as a Service	1.820.000 €
TIME	Translation Research Training: An integrated and intersectoral model for Europe	1.228.978 €
transLectures	Transcription and Translation of Video Lectures	3.125.000 €
TrendMiner	Large-scale, Cross-lingual Trend Mining and Summarisation of Real-time Media Streams	3.272.000 €
TTC	Terminology extraction, translation tools and comparable corpora	2.025.000 €
X-Like	Cross-lingual Knowledge Extraction	3.550.000 €
xLiMe	xLiMe - crossLingual crossMedia knowledge extraction	2.987.000 €
TOTAL		160.898.710 €

Source: Compiled by the authors based in data provided by the EC

Table 28: HLT related projects H2020

Acronym	Title	Maximum EC contribution
Cracker	Cracking the Language Barrier	999.995 €
FREME	Open framework of e-services for multilingual and semantic enrichment of digital content	3.212.626 €
HimL	Health in my Language	2.949.571 €
LT-Observatory	Observatory for LR and MT in Europe	982.563 €
MixedEmotions	Social Semantic Emotion Analysis for Innovative Multilingual Big Data Analytics Markets	3.036.910 €
MMT	Modern Machine Translation	2.994.700 €
QT21	Quality Translation 21	3.977.428 €
TraMOOC	Translation for Massive open online courses	3.081.148 €
TOTAL		21.234.941 €

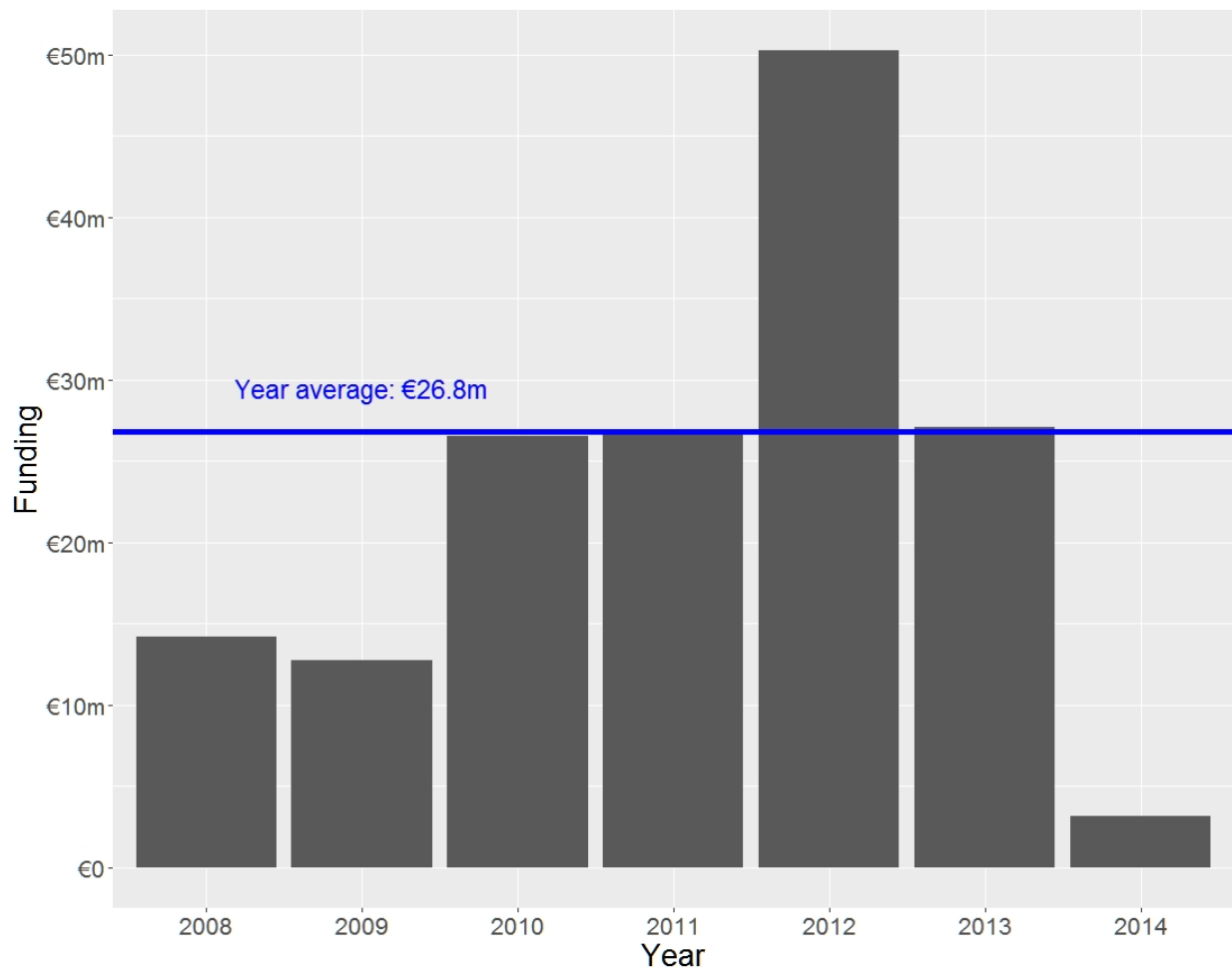
Source: Compiled by the authors based in data provided by the EC

Table 29: HLT related projects ICT-PSP

Acronym	Title	Maximum EC contribution
ATLAS	Applied Technology for Language-Aided CMS	1.486.810 €
BOLOGNA	Bologna Translation Service	1.580.000 €
CESAR	CEntral and South-east europeAn Resources	2.080.000 €
EASTIN-CL	Crosslingual and multimodal Search in a Portal for Support of Assisted Living	1.075.000 €
FLAVIUS	Foreign LAnguage Versions of Internet and User generated Sites	1.850.000 €
GALATEAS	Generalized Analysis of Logs for Automatic Translation and Episodic Analysis of Searches	1.850.000 €
iTranslate4EU	Internet Translators for all European Languages	1.968.000 €
LetsMT!	Platform for Online Sharing of Training Data and Building User Tailored MT	1.670.000 €
LISE	Legal Language Interoperability Services	1.250.000 €
METANET4U	Enhancing the European Linguistic Infrastructure	2.650.000 €
METANORD	Baltic and Nordic Parts of the European Open Linguistic Infrastructure	2.250.000 €
MORMED	Multilingual Organic Information Management in the Medical Domain	1.111.000 €
MultiLingualWeb	Advancing the Multilingual Web, Thematic Network	414.000 €
Organic.Lingua	Demonstrating the potential of a multilingual Web portal for Sustainable Agricultural & Environmental Education	1.750.000 €
PLuTO	Patent Language Translations Online	2.180.214 €
PROMISLingua	PeRformance Operational and Multilingual Interactive Services to support Compliance for SMEs in Europe	2.640.000 €
SUMAT	An Online Service for Subtitling by MACHine Translation	1.800.000 €
TOTAL		29.605.024 €

Source: Compiled by the authors based in data provided by the EC

Using the starting date of the project we have calculated the annual funding for HLT projects of the 7th FP for the period 2008 to 2014. The results are the following:

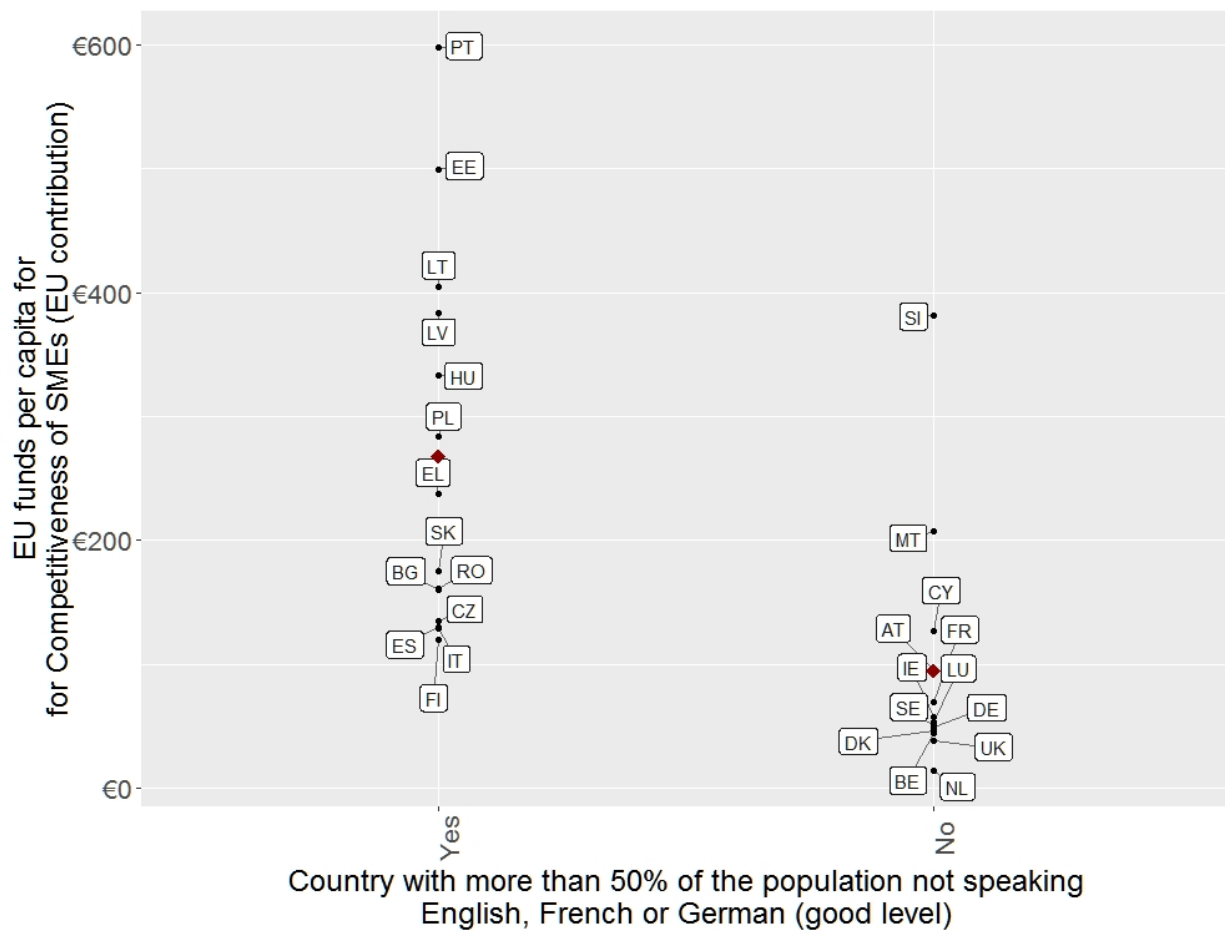
Figure 59: Annual contribution of HLT related projects FP7

Source: Compiled by the authors based in (European Commission, 2015h)

7.8 Allocation of European structural and investment funds (2014-2020)

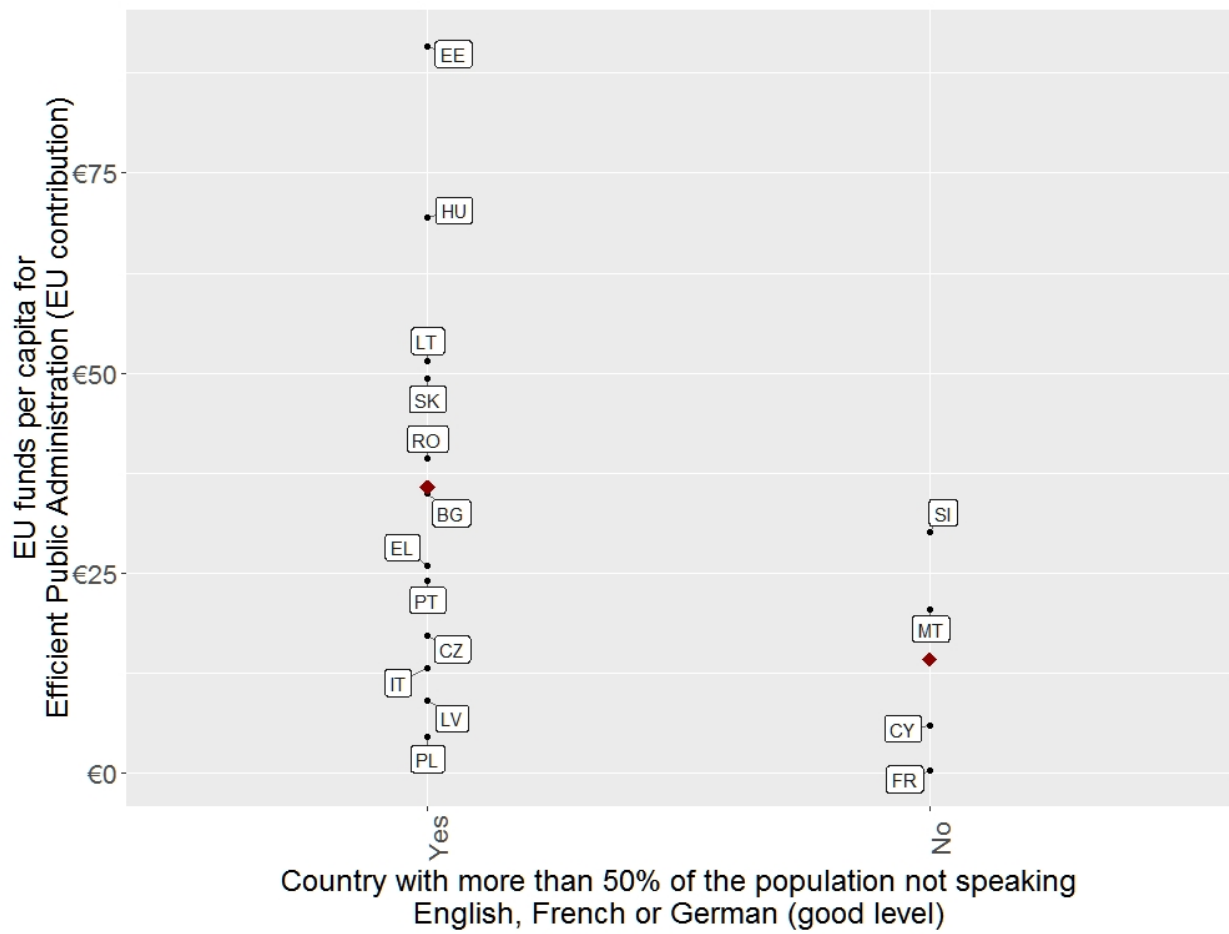
Figure 60 and Figure 61 show the contribution per capita of the EU per country for the 2014-2020 European Structural and Investment Funds regarding the themes “Competitiveness of SMEs” and “Efficient Public Administration” respectively. We classify the countries depending on having more than 50% of the population not being able to speak English, French, or German at a level at least good. Those countries that receive a higher amount of European Funds show also higher language barriers with the three majority European languages, and therefore are more likely to be benefited by using HLT.

Figure 60: Per capita EU fund contribution by country for Competitiveness of SMEs (depending on language barriers with majority languages)



Source: Compiled by the authors based on (European Commission, 2016f)

Figure 61: Per capita EU fund contribution by country for Efficient Public Administration (depending on language barriers with majority languages)



Source: Compiled by the authors based on (European Commission, 2016f)

The EU is a unique endeavour involving more than 500 million citizens sharing about 80 different languages, and while multilingualism is a key feature, it is also one of the most substantial challenges for the creation of a truly integrated EU. Language barriers have a profound effect on cross-border public services, on fostering a common European identity, on workers' mobility, and on cross-border e-commerce and trade, in the context of a Digital Single Market. The emergence of new technological approaches, based on increased computational power and access to sizeable amounts of data, are making Human Language Technologies (HLT) a real solution to overcoming language barriers. However, several challenges, such as market fragmentation and unsubstantial and uncoordinated funding strategies, are hindering the European HLT community, including research and industry.

This is a publication of the Scientific Foresight Unit (STOA)
EPRS | European Parliamentary Research Service, European Parliament



PE 581.621
ISBN 978-92-846-0698-6
doi: 10.2861/136527
QA-02-17-247-EN-N

This document is prepared for, and addressed to, the Members and staff of the European Parliament as background material to assist them in their parliamentary work. The content of the document is the sole responsibility of its author(s) and any opinions expressed herein should not be taken to represent an official position of the Parliament.