BRIEFING

European Parliament

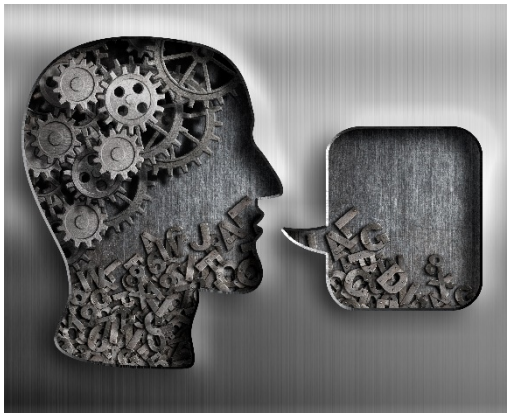
European Day of Languages: Digital survival of lesser-used languages

SUMMARY

Since 2001, Europe has marked European Day of Languages each year on 26 September, in order to focus attention on its rich linguistic diversity. The European Union boasts 24 official languages, and around 60 regional and minority languages are spoken across the Member States. Europe's linguistic mosaic also includes a variety of sign languages spoken by half a million people, heritage languages such as ancient Greek and Latin, as well as Esperanto – a planned international language created in Europe.

According to the United Nations Educational, Scientific and Cultural Organization (Unesco), many world languages, including European ones, are endangered and could disappear due to the dominant role of languages such as English with a huge population of native speakers and other learners. Regional and minority languages (RMLs) together with smaller state languages – the 'lesser-used languages' – are under serious threat of extinction.

This threat is exacerbated by digital technology. The future of RMLs depends to some extent on their presence in new digital media. Young people communicate and seek information mainly via the internet. If online content is only available in dominant languages, lesser-used languages could become 'digitally extinct'. However, digital technology is not necessarily a death sentence; it can also offer a rescue kit. Online education, online language learning and language technologies can help revitalise endangered languages. To achieve this objective, huge efforts are needed by speakers' communities and language technology specialists to gather data, analyse and process it, in order to create language tools. With such tools, young people can create content in lesser-used languages and expand their use.



In this Briefing

EPRS | European Parliamentary Research Service

Author: Magdalena Pasikowska-Schnass
Members' Research Service

PE 652.086 – September 2020

EN

Background

Since 2001 the European Day of Languages has been devoted to linguistic diversity across the continent, and beyond. The Council of Europe and the European Commission have joined forces to promote language learning and multilingualism, as well as to protect Europe's linguistic heritage by encouraging openness to different languages and cultures. Language learning addresses the issue of multilingualism, but for a long time now, the European Day of Languages has not focused on how to define linguistic heritage and how it connects with language learning.

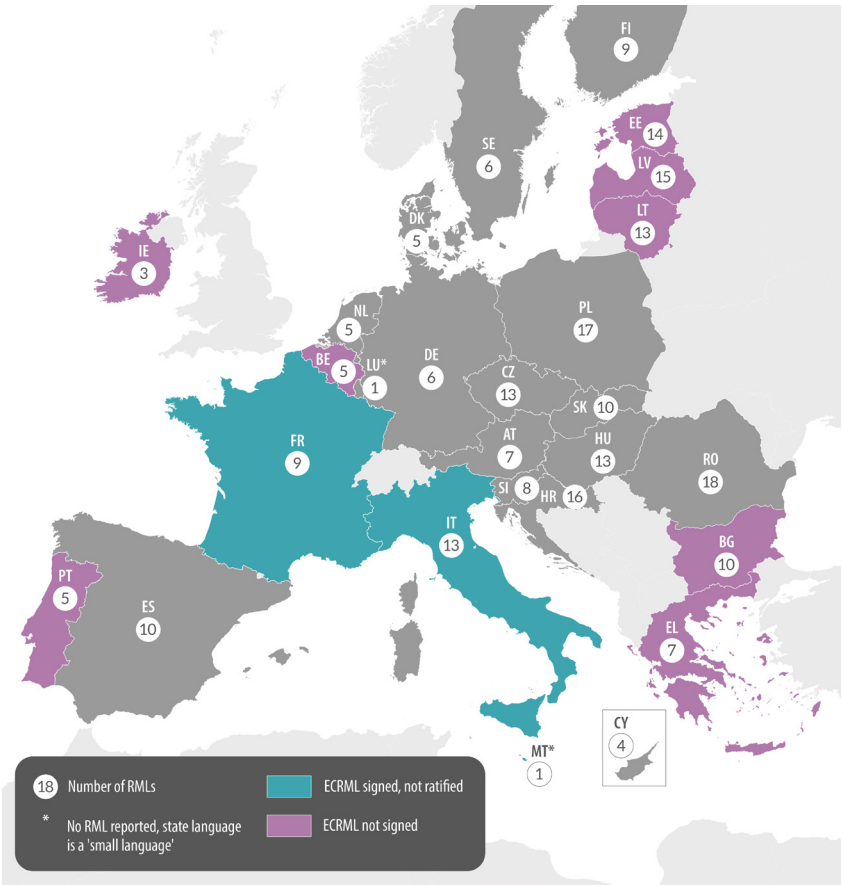
The European Union and linguistic diversity

The EU has no specific competence concerning the languages of its Member States. On the basis of Article 165(2) of the Treaty on the Functioning of the European Union (TFEU), dealing with education, it can support actions asserting the European dimension in education through teaching and dissemination of the Member States' languages. Respect for national and regional diversity is enshrined in Article 167 TFEU. Article 3(3) of the Treaty on European Union (TEU) stipulates that the EU shall respect its rich linguistic diversity confirming that diversity refers also to regional languages. Thus 'linguistic heritage' could include regional languages, but what about Latin and ancient Greek? Besides, there are still more aspects of linguistic heritage: minority languages, Esperanto – a planned international language based on European languages – sign languages, and dialects.

This diversity is very precious and needs to be protected, since Europe is one of the most linguistically homogenous continents: its population, which accounts for 7.1 % of humanity, speaks only 3 % of the world's estimated 6 000 languages.

While more than 1 600 'native tongues' are spoken in India hosting 16 % of the world population, an estimated 40 to 50 million people in the EU between them speak 60 regional and minority languages. Central and eastern EU states report a greater number of RMLs, reaching 18 in Romania, 17 in Poland and 16 in Croatia (see Figure 1). In most EU Member States in this area, RMLs are state languages of their neighbours like German in Poland and Czechia, or Hungarian in Slovakia and Romania, while in the Netherlands, Germany, Sweden, Denmark, Spain and France, non-state RMLs dominate (see Figure 2). The region is also home to some languages with very few speakers scattered among a number of Member States and in third countries, but these languages themselves are not state languages in any of the countries. The

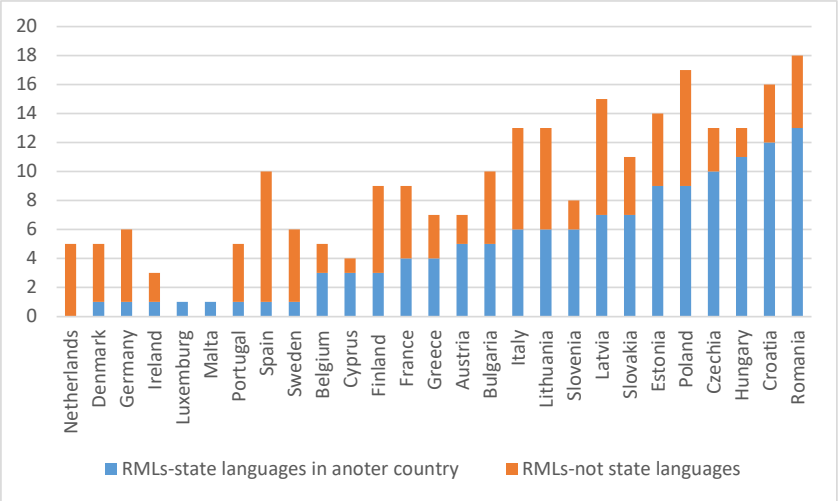
Figure 1 – Regional and minority languages in the EU



Source: Language diversity project, European Commission, 2012. Graphic: Lucille Killmayer, EPRS.

population of native speakers of Tatar ranges from 24 000 in Romania, to 5 000 in Poland and Lithuania and just 900 in Finland, while just 200 people in Poland and 3 000 in Lithuania speak Karaim. The EU is also home to Romani and Yiddish, non-territorial languages, which cannot be defined in terms of a particular region of the country but are spoken in various areas of many countries all over the world.

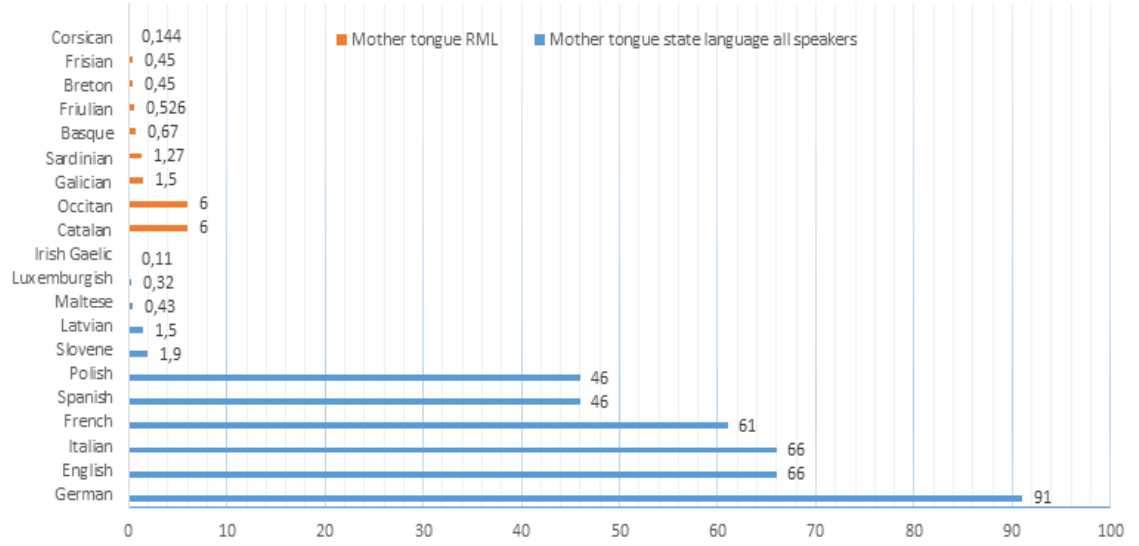
Figure 2 – RMLs as state and non-state languages in numbers



Source: Language diversity project, European Commission, 2012.

Three EU Member States stand out: the Netherlands which does not report any state language as RMLs, and Malta and Luxembourg – two Member States with very small populations, a condition which puts their state language in a situation similar to that of a minority language. Moreover, the three Member States figure among those with the highest levels of multilingualism, ranging from 95 % in Luxembourg to 86 % in the Netherlands. Ireland experiences a similar situation, as Irish, a state language, has a small population of native speakers and enjoys rather the status of a RML. Some RMLs' native speakers outnumber speakers of certain state languages and also qualify as lesser-used languages (see Figure 3).

Figure 3 – Populations of speakers of state languages and major RMLs in the EU



Source: Language diversity project, European Commission, 2012.

Protecting the EU's linguistic diversity

In 1992, the Council of Europe adopted the **European Charter for Regional and Minority Languages** (ECRML), which entered into force in 1998. The Charter focuses on the need to protect Europe's rich linguistic legacy, in particular its traditional RMLs, some of which are in danger of extinction if they are not protected and promoted. Not all EU Member States adopted the Charter (see Figure 1): 16 of the 27 Member States have signed and ratified it. Belgium, Estonia, Latvia, Lithuania, Greece, Ireland, Portugal and Bulgaria have not yet signed it. France, Italy, and Malta have not ratified it yet, and thus while they are committed to respecting their RMLs, they have not chosen any specific measures to promote them in education and to promote their use in public life, including in the media, social life and cultural activities. The Charter stresses that it protects languages and explicitly excludes any claims outside this scope. It also encourages states and other parties to cooperate in order to promote the RML they share, such as Basque – an RML in both Spain and France.

ECRML definition of RMLs

Regional and minority languages are traditionally used within a given territory of a state by a numerically smaller number of speakers than the rest of the population which speaks a different language. They do not include dialects or migrants' languages

Source: European Charter for Regional and Minority Languages.

There are four categories of RML; their status influences their fate:

- cross-border – a state language in one state is a minority language on the other side of the border (German in Poland, Polish in Slovakia, etc.);
- autochthonous cross-border – spoken in more than one state without having a state-language status: North Sami spoken in Sweden and Finland;
- autochthonous languages – languages without state-language status (like Breton in France); and
- non-territorial languages, which do not belong to one specific territory or state (such as Romani or Yiddish).

Cross-border RML are not particularly threatened thanks to their official status in the neighbouring or a third country. It is estimated that following the 2004 enlargement of the EU, 75 % of RMLs in the EU are cross-border ones. Non-territorial RMLs are dispersed among various Member States and can lack state support which weakens their position, particularly if they are purely oral ones, like Romani. These languages need some degree of standardisation to be linguistically processed. The status of an RML can vary in different EU Member States, and this too impacts their fate: Basque, together with Catalan and Galician, enjoys an official status in Spain, but is not officially recognised in France, and its promotion may differ accordingly, while Unesco considers this unique European language vulnerable.

Many RMLs are endangered, or considered vulnerable. Kashubian, Scots, Breton and Sami languages are classified as severely endangered, while Livonian and Cornish are critically endangered. For example, the Unesco Atlas of Endangered Languages reports 31 endangered languages in Italy, with four of them severely endangered, among them Gardiol having only 340 speakers in 2007 and Töitschu 200. Similarly, 26 languages are endangered in France, half of them severely.

Among factors of language vitality, an important aspect of RMLs, Unesco lists:

- language transmission among generations
- absolute number of speakers and their proportion to the total population
- shifts in domains of use: from universal to occasional use, as in rituals and ceremonies
- response to new domains, like new media and life experience
- availability of teaching material.

The advent of digital technologies could also have an impact on transmission and the number of speakers, depending on digital teaching material and digital presence.

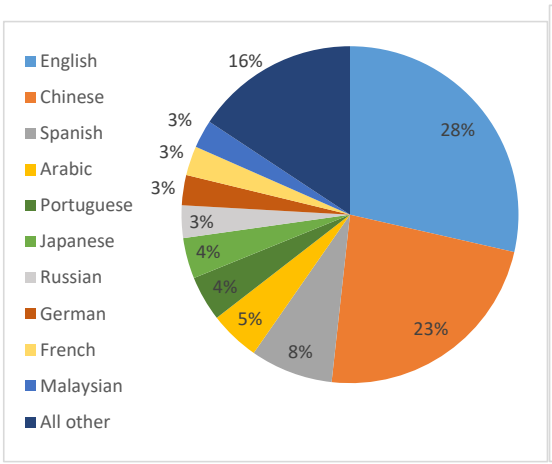
Protecting linguistic diversity in digital environment

A 2003 Unesco recommendation to promote online multilingualism and universal access to cyberspace calls for the development of multilingual and public domain content. It also urges access to networks in indigenous languages, in an effort to foster universal, equitable and affordable access to information, and the preservation of cultural and language diversity.

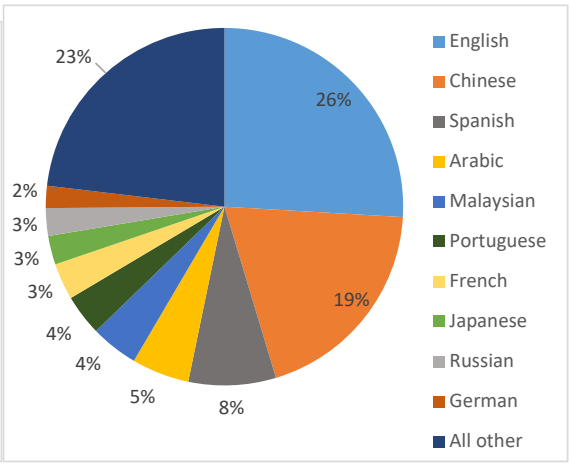
With 300 000 speakers being the estimated critical threshold for the survival of a language, ICT can play a vital role in RML promotion; but can also be a threat. RML presence on the Internet is referred to as 'digital survival' which is a particular challenge to lesser-used languages. According to a 2019 report by a group of ECRML experts, in the long run the future of RMLs depends on their presence in new media and digital mass communication. The report points to an urgent need for more sustainable RML media content targeted at young audiences, winning their engagement thanks to frequency and consistency. The authors also note that searching to retrieve information, news and content in an RML is an act of choice requiring management of cookies and online profiling. Moreover, exposure to media for bilingual or interested users of an RML is less likely to occur, as RML media become more marginalised in the general media environment.

A 2014 study on the use of Frisian by teenagers shows that it depends on the level of formality of the media. The more formal it is, the less Frisian is used (Facebook and Twitter 30 %, e-mails 15 %) while in less formal media like WhatsApp half of users write in Frisian. Among adults the proportions are higher but the trend is similar. The authors conclude that a number of RMLs are more vivid in their oral form, while the use of their written form is more problematic. The perception of an RML by its native speakers and the surrounding community is an important factor in the use of RMLs. Language technology (LT) support such as spell-checkers could encourage the use of RMLs and lesser-used languages. The presence of RMLs on the web depends on their inclusion in ICT. However, LT originates mostly in English-speaking countries, and mostly focuses on dominant international languages. The gap between them and RLMs in the digital environment is huge and widening as the internet is the main source of communication and knowledge for young generations.

Figure 4 – Languages used on the internet by share of internet users in 2013 and 2020



Source: The Internet World Stats (In Digital Language Divide), 2013.



Source: Internet World Stats, 2020.

A 2000 study noted that the major threat to linguistic diversity on the Internet is a multilingualism that is limited to a handful of major world languages, supported by machine translation, to the detriment of the great majority of smaller languages. The real danger comes from a facade of linguistic diversity that includes some dominant languages but excludes all others. Figure 4 confirms this trend, with minor changes in percentages among the same ten dominant languages. However,

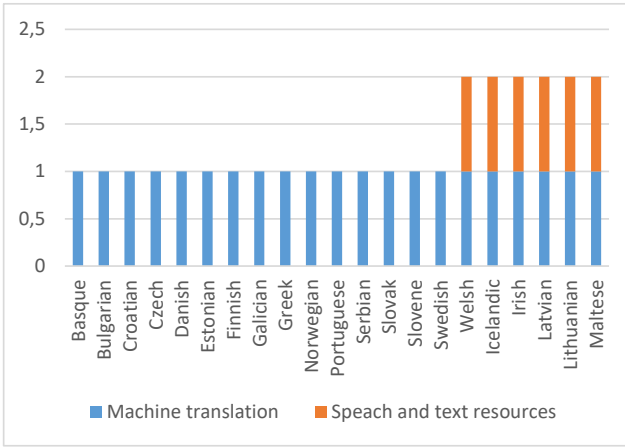
while the use in cyberspace of English and Chinese diminished slightly, the share of 'other languages' grew, but this composite figure does not give information about the possible marginalisation of certain lesser-used languages.

The European Roadmap for Linguistic Diversity proposed in 2015 by the European Network to Promote Linguistic Diversity stressed the importance of RMLs, not only as communication tools, but also as an expression of values and aspirations, resulting in the need to develop ICT tools for them.

Most European languages risk digital extinction if there is insufficient LT support for them, which is the case for the majority of European languages. ICT can be particularly challenging for languages that are under-represented on the internet, but it can also be beneficial, as language technologies and data computing offer solutions for language learning and translation.

Such a situation requires the development of LT for all European languages, particularly RMLs, as well as languages that depend on their oral tradition and lesser-used languages, including state languages like Luxemburgish, Maltese, Irish, and Estonian. Smaller, sign and oral languages lack the full range of language resources such as linguistic corpora and electronic dictionaries. Data needed to develop human language technologies such as machine translation, and speech and text resources are scarce or limited (see Figure 5).

Figure 5 – Coverage with machine translation and speech and text resources for European languages



Source: Key Results and Cross-Language Comparison, META-NET White Paper, data 2012-2014.

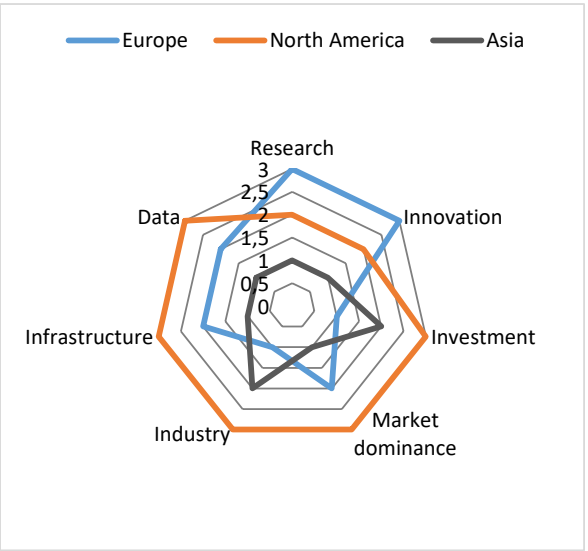
European Parliament

Already in its 2013 resolution on endangered European languages and linguistic diversity in the European Union, the European Parliament noted that digitisation could be a way of preserving endangered languages and called on local authorities to gather their online books and recordings as a resource to prevent these languages' extinction. It also called on the European Commission to involve younger generations in initiatives on digital media in an effort to revitalise endangered languages.

Five years later, Parliament adopted a resolution on language equality in the digital age. It followed a 2017 study commissioned by Parliament 'Language equality in the digital age. Towards a Human Language Project'. The study focuses on language barriers in a linguistically fragmented European Single Market and in particular in the European Digital Market and in European language technologies.

Parliament's resolution recognises that 24 official EU languages, more than 60 RMLs together with sign languages and migrant languages, is a mosaic of languages which is a challenge to a multilingual Digital Single Market, particularly as

Figure 6 – Position of European machine translation market vs North America and Asia



Source: Final study report on CEF, Automated Translation value, European Commission, 2019.

regards digital technologies where the EU lags behind major technology players such as North America and Asia (see Figure 6).

The situation further hinders certain communities (elderly, disadvantaged, less educated populations and speakers of lesser-used languages). The resolution called on the Commission to create a centre for linguistic diversity, and define minimum language resources for all European languages, such as data sets, lexicons, speech records, translation memories, and encyclopaedic content. It pointed to the need for an action plan to promote linguistic diversity and overcome language barriers in the digital area within a reviewed Framework Strategy for Multilingualism as a step towards an EU policy on language technologies, particularly for RMLs and lesser-used languages. Such a strategy would address shortages in coordination of research, development and innovation, and insufficient funding.

The EU and language technology

A Strategic Agenda for the Multilingual Digital Single Market drafted at the 2015 Riga summit defined a programme consisting of innovative technology solutions for businesses and public services, LT services, platforms and infrastructure, and priority research themes.

To address the challenges highlighted above, the European Commission has supported research and innovation projects through funding from the Horizon 2020 research programme, Connecting Europe Facility and the Erasmus + programme. A number of recent initiatives have paved the way for bringing together communities that are fragmented according to the language they speak. Among these is Horizon 2020 funding for the European Language Resource Coordination network and data project, and other initiatives such as CRACKER – Cracking the Language Barrier Federation, working on coordination, evaluation and resources for European machine translation research, the Multilingual Europe Technology Alliance network META-NET, involved in LT research and engineering, and the Common Language Resources and Technology Infrastructure CLARIN projects, with languages such as Basque, Breton, Welsh and Frisian, making digital language resources available to scholars, researchers and other interested parties.

Erasmus+ funded the Digital Language Diversity Project which aims at enhancing the digital presence of lesser-used and endangered languages, focusing on Breton (France), Basque (Spain), Karelian (Finland) and Sardinian (Italy). It also worked on a training programme for RML speakers to help them produce digital content and learning materials in their respective languages. It aimed at identifying challenges RML users face related to the digital presence of their languages and at guiding them via the digital language vitality scale. Recommendations in the 'digital language survival kit' help language communities to self-assess the digital fitness of their RML against a range of criteria such as the presence of digital infrastructure and skills, connectivity, localised social networks, operating systems and software, machine translation, and dedicated internet domain names. Finally, a roadmap to digital language diversity, a tool for decision-makers to help them choose an appropriate technological solution for the use of a given language on a digital device, offers policy recommendations for example for state administrations to develop digital services in lesser-used languages. As a result, the participants in the project hope to provide the necessary conditions for software developers, SMEs and industry to manage to provide products and services using RMLs such as

Sign languages and ICT

ICT would be helpful for processing sign languages to facilitate communication among various sign language communities as well as with communities using oral communication. The latter face the difficulty of transition between a three-dimensional space playing a grammatical role since spatial organisation of gestures, mimics and body language play a role in communication, and the linear nature of oral languages where relation between words playing various grammatical roles is defined by specific pronouns which do not exist in sign languages. This factor is of concern for both oral and written forms of languages. The lack of multilingual data on sign languages is a significant barrier for language technologies researchers to progress.

Source: Language equality in the digital age, study, EPRS, EP, 2017.

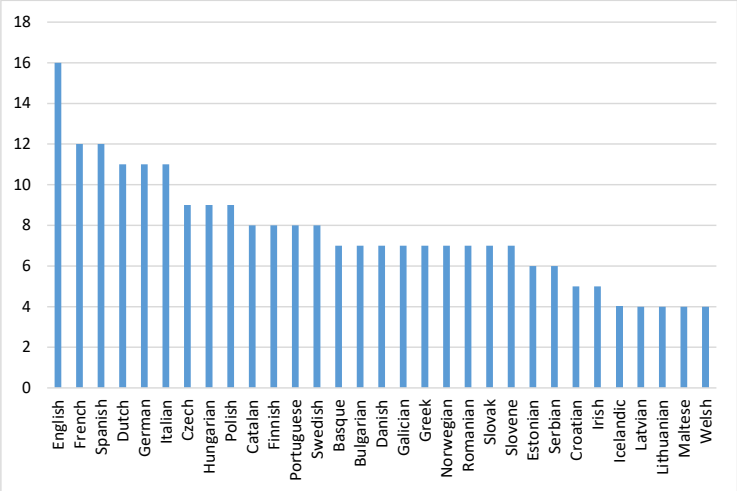
subtitling, localised interfaces for social media platforms, spelling correctors and keyboards and video games. Twin initiatives allow language communities to keep in touch and develop their linguistic resources. The research results gathered in recommendations point to the fact that digital development of lesser-used languages depends on the linguistic community actions and engagement and should not rely on industry, which is not interested in small and complex languages.

This conclusion is confirmed by a 2019 study on LT conducted within the framework of the Connecting Europe Facility programme. The programme is devoted inter alia to the investigation of digital infrastructure opportunities and barriers the EU faces while competing with North America and Asia in this domain. However, the study did not receive responses from RML users and institutions, except for users of Basque and Ladino, even though LT are essential for the digital presence of languages, particularly for lesser-used languages. The study concludes that industry in the EU focuses on the main languages: English, German and French, even though there is a strong experience in the small languages present on its market. However, this expertise offers few business opportunities due to the limited scope of the market related to such languages and the low availability of data about them. Nevertheless, the report stresses the EU's commitment to protect its linguistic and cultural diversity. It also draws attention to the inclusivity which public services should take into account while dealing with lesser-used languages. The authors point to the dominant position of Europe in research and innovation, for example in machine translation, countered by a very weak position in investments and industrial use of the results of the research and innovation, which again corroborates the conclusion of the Digital Language Diversity Project (see Figure 6).

A research project 'European Language Grid' (ELG) funded by the research programme, Horizon 2020 aims at addressing this complex situation for LT industries in the EU, characterised by linguistic fragmentation and a high presence of SMEs on the LT market. Some of them have gained world-class expertise, but face huge competition from dominant giants, including in their small national markets. In order to make the Digital Single Market function in a coherent way, this linguistic fragmentation must be addressed with respect to lesser-used languages in the EU and beyond. The grid, under development, will deploy a scalable cloud platform providing access to commercial and non-commercial LT for all European languages, together with data sets and resources. It will result in a browsable, searchable and explorable online catalogue that companies, academic organisations and individual researchers could filter and search for domains, sectors, regions, countries, languages, service types, data sets, etc. The first release of the ELG, coordinated by 32 National Competence Centres across Europe and the European Language Technology Board, took place in May 2020. After the final third release in late 2021 it should include more than 800 functional LT services and more than 3 500 LT data sets, corpora, resources and models.

Another Horizon 2020 project, META-NET, gathered data on LT support such as machine translation, speech processing, text analysis, and speech and text resources, for a range of languages, including lesser-used languages in both state and stateless frameworks. These data are essential for language processing and its online

Figure 7 – LT coverage of European languages



Source: Key Results and Cross-Language Comparison, META-NET White Paper, data 2012-2014.

presence. The results show (see Figure 7) that not even English scores 'excellent' (a score of 16 or more) in LT support. Four lesser-used languages, all state languages, have weak or no support (score 4 to 6). A majority of languages have moderate or fragmentary support (8-11). Moreover, 21 out of 31 investigated languages are not supported by machine translation, and six of them (including five state languages) have weak or no speech and text resources (see Figure 5). This conclusion gives us a picture of the real danger of digital extinction for an estimated 80 % of European languages.

Importantly, the European Commission currently provides European SMEs, public administration and citizens with free of charge secure automated translation, eTranslation services in more than its 24 official languages (also Icelandic and Norwegian), a number of which belong to lesser-used languages at serious risk of digital extinction.

In order to enhance digital linguistic diversity it is also important to promote RMLs among young people. Erasmus+ educational projects include eTwinning to promote online cooperation even in lesser-used languages thanks to projects such as 'Celtic cousins' between Ireland and France on Celtic culture. Joint initiatives aim at raising awareness of lesser-used languages in schools through partnership projects, staff mobility and youth exchange programmes. Many of these projects have been digitalised or use ICT tools and are available online.

The way forward

Multilingualism, which also covers RML as a policy area under the current European Commission, is under the responsibility of the 'Accessibility, multilingualism and safer internet' unit as a part of the European Digital Strategy working on policy, research, innovation and deployment of key enabling digital language technologies and services for both consumers and businesses. This is a clear sign of the recognition of the importance of linguistic diversity and the use of digital technologies for its preservation. In the next Multiannual Financial Framework (2021-2027) LT will continue to be covered by funding from the Digital Europe programme for capacity-building (artificial intelligence, data space, high power computing), deployment, best usage (cloud infrastructures, deployment of modern public services) and amplifying the best use of digital technologies (LT). The next research programme 'Horizon Europe' under the pillar 'Digital Humanities and Language Technologies' will support language preservation by collecting language resources to prevent digital extinction which endangers most European languages. Its Next Generation Internet research programme will support research on digital language transparency.

eprs@ep.europa.eu (contact)

www.eprs.ep.parl.union.eu (intranet)

www.europarl.europa.eu/thinktank (internet)

http://epthinktank.eu (blog)