Enabling additional official languages in the EU for 2025 with language-centred AI

Kepa Sarasola and Nora Aranberri
HiTZ Basque Center for Language Technologies,Ixa NLP group
University of the Basque Country,
kepa.sarasola@ehu.eus

Keywords: *Machine translation, Artificial Inteligence, European official language, Under-resourced languages*

Abstract

Diversity is socially beneficial, and therefore, language diversity too. Then, why not grant the official status to more languages at European level? In 2021, the costs associated with it are no longer an excuse. The potential of the new generation of machine translation, voice synthesis and voice recognition applications, which are getting readily available even for under-resourced languages like Basque, will allow Europe to promote and integrate more and more of its languages, and with it, its citizens.

Introduction

We are in times of change. A technological revolution is taking place right here and right now. The quality achieved by machine translation applications, voice synthesis and voice recognition today leads us to predict that there will be substantial changes in social communication and in the areas of use of each language in just a few years.

The futuristic scenario set out in the novel "Qui vol el Panglòs" written by Antoni Olivé 30 years ago is close to becoming a reality. The novel is a reflection on what could happen in a society with an unstable linguistic balance (like Catalonia) if a device (Panglòs) were invented that would allow people to understand any language and to communicate with everyone else only using their own language. In 2021, Panglos is not a reality. There is still a margin of error when using machine translation, which is bigger or smaller depending on the language pair involved. Consequently, human post-editing is still necessary, but human translators tend to achieve a much higher efficiency when using this tool. However, it is highly likely that the flow of knowledge will be smoother between cultures and among the scientific community, and as a result the pace of social and scientific progress will increase dramatically in the short term.

What is important to note at this stage is that these technologies are not only applied to widely spoken languages; recently, research is also being conducted for low-resourced languages with promising results. Taking Basque as example, in the following lines we aim to (1) present instances where new technological tools are proving beneficial for under-resourced languages; (2) describe current research and development projects that afford us a glimpse of the next scientific breakthroughs; and (3) put forward ideas to take full advantage of the technological advances which would enable the promotion of under-resourced languages in Europe.

The experience we describe here demonstrates that the new technological developments can be used to promote and consolidate the use of under-resourced languages across Europe. Moreover, it

shows that these technologies would allow for a dynamic and affordable provision of language services that could vary according to the demand, and this would make it possible to recognise all languages as official.

1. Quality leap in language technology

Natural Language Processing (NLP) and Machine Translation (MT) technologies, which are part of Artificial Intelligence, are in the heart of today's information processing software. These exploit large amounts of structured and unstructured data collected from texts, as well as from websites and social networks. NLP and MT processors make it possible to analyse and exploit texts in domains such as education, health, law, tourism or the economic market. Language processors currently offer a long list of applications such as language checkers, language learning systems, machine translation, intelligent search, named entity detection (proper names, dates, brands, products, etc.), document classification, document clustering, document filtering, automatic summary creation, data extraction from documents (sentiment analysis and opinion mining), reputation tracking and monitoring in social networks, alert generation, document queries, chatbots, etc.[1]

1.1. Language technology as a tool for language revitalisation

Since 1968 Basque has been inmersed in a revitalisation process facing formidable obstacles. However significant progress has been made in numerous areas. Six are the main factors identified to explain its relative success: 1) the implementation and acceptance of the Unified Basque (Batua), 2) integration of Basque in the education system, 3) creation of media in Basque, (radio, newspapers, and television); 4) the established new legal framework, 5) collaboration between public institutions and people's organisations, and 6) campaigns for Basque language literacy(Agirrezabal, 2010). While those six factors influenced the revitalisation process, the extensive development and use of language technologies is also considered a significant additional factor (Alegria & Sarasola, 2017).

Adding to the thorough work done for Basque since the early days of language technologies, it is expected that the great technological improvements that have taken place in the last six years, such as big data or neural networks, may bring about a second big leap in its recovery. This new upsurge may be qualitatively greater for our society. Significant improvements in speech processing, machine translation and text analysis could make a major contribution to facilitating the use of Basque. Needless to say, the technological development did not occur without effort. NLP research for Basque has been continuous and active, mainly stirred by a local research group (IXA group, University of the Basque Country). Efforts have been made to be up-to-date with the latest research trends and to disseminate the results obtained for Basque in key research conferences and international fora. Of course, this was supported with master- and doctoral-level programmes on NLP.

The Basque language is currently one of the pioneers in methodologies for the promotion of minority languages. This can be seen in the scientific congresses in the field or in the origin of the students who attend the x master's degrees offered by our university. If we make good use of the possibilities offered by technology, in the medium term our course will also be an international benchmark, with

1https://plantl.mineco.gob.es/tecnologias-lenguaje/PTL/Bibliotecaimpulsotecnologiaslenguaje/Detalle%20del%20Plan/Plan Plan for the Advancement of Language Technology. Mineco Spanish Government.

solutions not only for Basque, but also for different languages in international forums and in a multilingual market.

1.2. Rapid increase in LT development

Development in the field of language technology has been remarkable in the last six years. For example, the first scientific publication using deep learning in machine translation appeared in 2014, authored by Bahdanau, Cho and Bengio. One year later 90% of the MT systems winning research challenges were neural systems. The breakthrough in 2015 was the use of Attention based NMT systems. In 2017, further improvements were obtained from the use of the Transformer architecture in neural networks. While initial work was carried out for English given the resources it avails for experimentation, it took just one year to successfully implement a system for Basque, and by 2019, there were not one but five successful systems for Basque.

Currently, contributions are being made from the field of machine translation for languages with few resources. The research community is concerned with identifying techniques and strategies to work with languages with limited resources. For example, some efforts have focused on using multilingual data to enrich the tools' knowledge of the low-resourced languages (Fan et al., 2021). Similarly, it is only a key aim to build domain-specific tools that are also scalable for industry.

Translation technology has taken giant steps forward, opening new challenges to bring multilingual services to our society. We estimate that the production of translated documentation could be 10 times bigger in a few years without increasing the number of professional translators.

1.3. New powerful tools

Over the last three years, several language applications have emerged in the technological landscape of Basque that could prove extraordinary catalysts for promoting the use of the language in the public sphere. These applications allow speakers of other language to understand text and speech in Basque, and Basque speakers to understand texts in other languages. The following are some of the most noteworthy applications:

- ***Elia.eus***[2], ***itzuli+***[3], and ***batua.eus***[4] machine translators. Besides the well-known Google Translator, there are three locally developed neural systems that provide high quality translation proposals.
- **Content Translation** tool[5] is allows Wikipedia editors to create translations right next to the original article and automates the boring steps: copying text across browser tabs, looking for corresponding wiki-links, wiki-categories, wiki-templates and programmed components etc. The deep intrinsic multilingualism of Wikipedia and Wikidata allows, for example, easy translation for all languages of the infoboxes that appear top right in Wikipedia articles. Content Translation offers translation from/to Basque by using elia.eus, Google tranlate or

2 https://elia.eus/
3 https://www.euskadi.eus/itzuliplus/
4 https://www.batua.eus/
5 https://www.mediawiki.org/wiki/Content_translation

Yandex. A first international event[6] has been recently organized in 2021 to allow the research community to take stock of the progress made so far and to identify new avenues for future work.

- **Aditu**[7] bilingual speech recognition. The *Aditu* web service recognises both Basque and Spanish speech. It should also recognize English and other languages by 2021. It allows to obtain high quality instant transcriptions, automatic generation of subtitles, and direct transcription from the microphone. Anyone can use these applications and then correct transcriptions or subtitles on the online editing interface.
- **Interprest**[8] system for interpreting. The main goal of the system is to enable low-cost and portable interpretation services for different types of events. It is based on mobile phone communication systems, that is, it is a wireless system. The communication process is simple: the mobile phone of the interpreter sends the audio through a small microphone and each attendant can use his/her own phone to listen to the simultaneous translation. Interprest was a technological platform powered by "San Sebastián 2016," the European Capital of Culture.
- **Bidaide**[9] is a web service that allows the visitors of a museum, route or building to read or listen to descriptions and general information about the sites on their own mobile and in their own language (Cortes eta al, 2018). Visitors access the information in various ways: by scanning QR codes located in key areas, by GPS positioning (in outdoor routes), or by automatic Bluetooth proximity activation. This makes it accessible even for people with reduced or null vision. Additionally, this platform also offers the manager of the visited site advanced language resources to create the texts and audios in all the relevant languages: machine translation is used to translate texts, while speech synthesis is used to produce audio materials. For accessibility purposes Bidaide uses technologies for speech synthesis and speech recognition. Bidaide uses the visitor's location to guide him/her along outdoor routes by means of GPS, and indoors through various Bluetooth transmitter beacons. The multilingual content is stored online and it is managed by the site management team.

6 https://ctn.hkbu.edu.hk/wikiconf2021/ Understanding Wikipedia's Dark Matter. Translation and Multilingual Practice in the World's Largest Online Encyclopaedia
7 https://aditu.eus/
8 https://talaios.coop/2016/09/interprest/
9 http://bidaide.elhuyar.eus

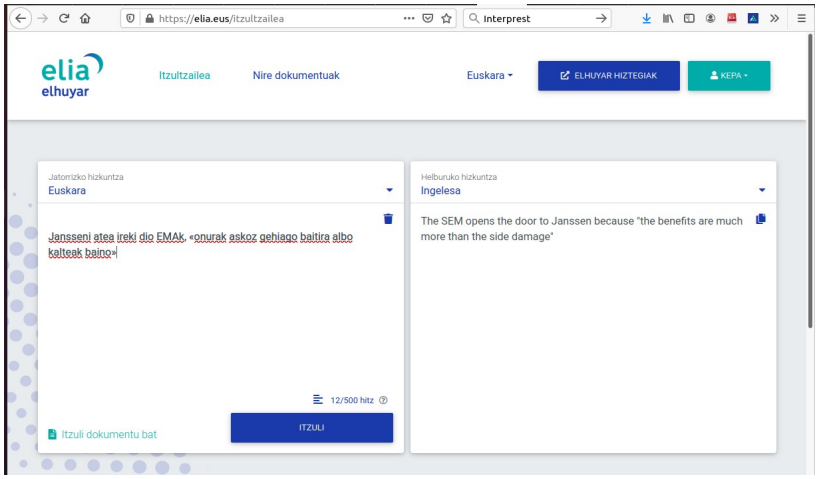

Fig. 1: elia.eus mchine translator.

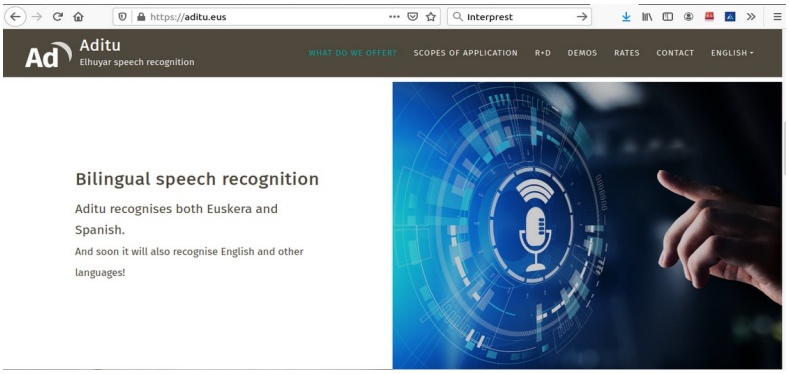

Fig. 2: Aditu bilingual speech recognition.

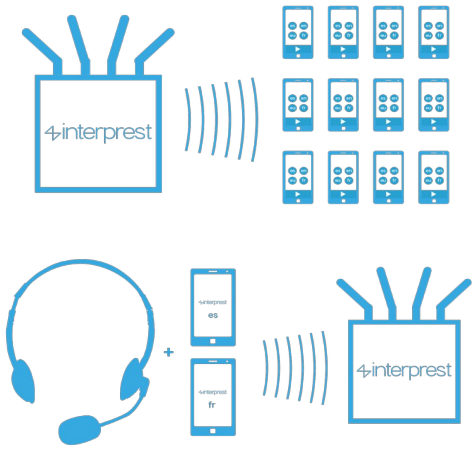

Fig. 3: Interprest system for interpreting.

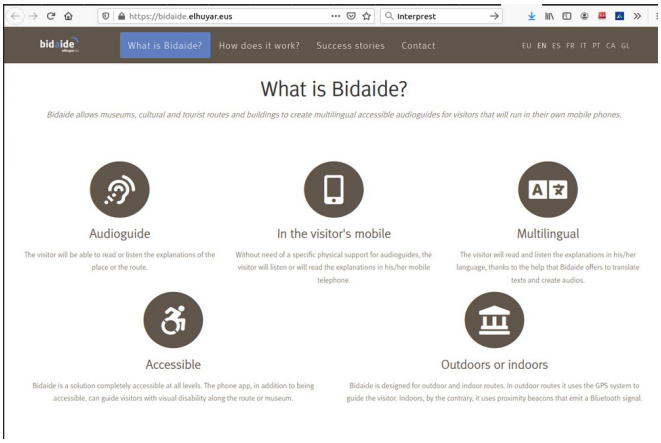

Fig. 3: Bidaide**,** a web service to that allows the visitors of a museum, route or building to read or listen to descriptions and general information about the sites on their own mobile and in their own language.

2. Projects for low-resourced languages

Participation in European initiatives that seek to advance in language technologies for low-resourced languages has also been key to the digitalization of Basque. We here briefly describe two such projects, which aim to enhance language technology applications within the European territory: Linguatec and the European Language Equality project.

2.1. Linguatec: cooperation among the languages of the Pyrenees

Linguatec is a project funded by FEDER via POCTEFA (INTERREG V-A Spain-France-Andorra programme). The main objective of Linguatec is to develop, test and disseminate new innovative linguistic resources, tools and solutions for a better digitalization level of the Aragonese, Basque and Occitan languages (Aldabe et al., 2019).[10]

The consortium consists of the following partners: 1) Elhuyar Fundazioa (working on Basque and Spanish languages); 2) Lo Congrès Permanent de la Lenga Occitana (Occitan and French languages); 3) the University of the Basque Country (Basque and Spanish languages); 4) CNRS Toulouse Delegation Regionale Midi-Pyrenees (Occitan and French languages); 5) Euskaltzaindia – Real Academia de la Lengua Vasca (Basque language); 6) Sociedad De Promoción y Gestión del Turismo Aragonés (Aragonese language).

In total, the project has developed thirteen main applications since 2018 that facilitate the cooperation and interoperability between the languages of the Pyrenees:

- A new translation system for Basque-French.

10 https://linguatec-poctefa.eu/eu/sarrera/#pll_switcher

- A new translation system for French-Occitan.
- A new translation system for Spanish-Aragonese.
- An improved neural translation system for the Spanish-Basque.
- A translation app for the languages of the Pyrenees: Basque-French, Basque-Spanish, French-Occitan and Spanish-Aragonese.
- VOTZ, the first tool for speech synthesis in Occitan.
- ReVOc, the first tool for speech recognition in Occitan.
- A Northern Basque speech recognition system.
- New monolingual and bilingual lexicons and morphosyntactic/syntactic analysers for Occitan.
- The Handbook of the Unified Basque: "Euskara Eskuz Esku Digitala" by Euskaltzaindia, the Academy of Basque.
- An on-line dictionary of Aragonese, and a roadmap for the Digitalization of Aragonese.
- A multilingual semantic search engine.
- A system for measuring the vitality of Occitan, Basque and Aragonese.

2.2. European Language Equality: a roadmap to a full Digital Language Equality in Europe by 2030

In September 2018, the European Parliament endorsed the report on Language Equality in the Digital Age presented by Jill Evans MEP of Wales with 592 MEPs voting in favour, and with only 45 against and 44 abstentions. Although the report did not become a law, it was a declaration made by the European Parliament which could be used as a reference by all European countries. Until that moment, there were no laws or declarations by the European Parliament to protect low-resourced languages, and therefore, decisions regarding their use and promotion remained in the hands of the local legislations of each country, who could easily disregard them. The report was a decisive step forward.

Among other things, the report states that "*The EP calls on the Commission and the Member States to develop strategies and policy action to facilitate multilingualism in the digital market; requests, in this context, that the Commission and the Member States define the minimum language resources that all European languages should possess, such as data sets, lexicons, speech records, translation memories, annotated corpora and encyclopaedic content, in order to prevent digital extinction*".

This should pave the way and boost local initiatives for the technological development of minority languages. It remains to be seen how this declaration materialises and to what extent real coverage is given to non-state languages in the coming years.

In this context, the primary goal of the European Language Equality (ELE) project is to prepare the European Language Equality Programme in the form of a strategic research, innovation and implementation agenda and a roadmap for achieving full Digital Language Equality (DLE) in Europe by 2030. [11]

The preparation of the plan to achieve DLE in Europe by 2030 requires:

- the accurate and up-to-date description of the 2021 state of technology support for Europe's languages,

11 https://libereurope.eu/project/european-language-equality-ele/ European Language Equality (ELE) Project

- the preliminary definition for achieving full Digital Language Equality in Europe by 2030, and
- the identification of gaps and issues regarding LTs, also considering neighbouring disciplines, particularly language-centric artificial intelligence.

The Consortium consists of a total of 53 members: 5 core partners, 9 networks, associations and initiatives, 9 companies and 30 research organisations. In addition to all official European languages, their expertise covers several unofficial, regional and minority languages, either through consortium partners or through the umbrella organisations ELEN and EC-SPM. The consortium as a whole brings together research and industry partners as well as wider networks representing a very broad range of stakeholders that have come together to achieve full DLE for all European languages.

3. Additional EU official languages for 2025?

Most countries in the world are multilingual and many are officially multilingual. Taiwan, Canada, Philippines, Belgium, Switzerland, and the European Union are examples of official multilingualism. Under their system, all government services are available in all their respective official languages. Also, each citizen may choose their preferred language when conducting business.

In the following paragraphs, we examine some features of practical multilingualism and legal multilingualism in Europe and put forward a number of ideas from the *Global Trends to 2035* report endorsed by European Parliament (EPRS, 2019).

3.1. Practical multilingualism: information in “many” languages

The implementation of practical multilingualism is relative. For example, the website of the European Commission states that the information is provided in the 24 official EU languages: Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish and Swedish. However, this is not always the case. Indeed, we can read this warning on the website of the Commission:[12]

> “*We aim to strike a reasonable balance between respect for speakers of the EU's many languages and practical considerations such as the cost of translation.*
>
> *Some types of content, such as legislation, are always available in all EU languages. Other content might be available in 1 language only or in a combination of languages that user research tells us will enable us to reach the largest audience in the most efficient manner.*
>
> *All content is published in at least English because research has shown that with English, we can reach around 90% of visitors to our sites in either their preferred foreign language or their native language. We also monitor users' behaviour to see when they are trying to view pages in a given language, so that we can request translation of the most 'in-demand' pages.”*

What is clear from the first paragraph is the criteria applied by the EU for practical multilingualism: “practical considerations such as the cost of translation”. This is, therefore, one of the

12 https://ec.europa.eu/info/language-policy_en European Comission. Language policy. Information in many languages

key aspects that low-resourced languages, and other main languages too, must address if all languages are to have the same status.

3.2. Legal multilingualism in Europe: 24 state languages, tertium genus languages and others

The territory of the European Union is made up of a rich and wide-ranging universe of languages, which is not always circumscribed to the State languages. The existence of multilingualism is one of Europe's defining characteristics and it should remain so in the constantly evolving model of Europe's political structure. Nonetheless, until now, the official use of languages has been limited to the State languages and has been based on a concept of state monolingualism that has led to a first level of hierarchization among the languages of Europe. This has shaped the very concept of European language diversity (Urrutia, 2005).

The legal scope for the recognition of European language diversity referred to in Article II-82 of the Treaty establishing a Constitution for Europe ("The Union shall respect cultural, religious and linguistic diversity.")[13] and the possible measures to implement the precept that might constitute the definition of a true European language policy on regional or minority languages is yet to be defined. The Treaty of the European Constitution contains various language-related references that can be grouped in two main categories: firstly, references to the constitutional status of languages, and secondly, references to the recognition of European language diversity.

The constitutional language regime is based on the concept of Constitutional languages, but the official status of languages is not governed by this rule. A second level of constitutional recognition of languages is introduced, which refers to official languages in the member states (Catalan, Basque, Galician, etc.). These languages, however, are excluded from the right to petition; they constitute a tertium genus, an intermediate category between the languages benefiting from the language rights recognized under the Constitution and those for which no status is recognized in the European institutional context. The legal use of the second, intermediate category depends on the development of standards, i.e., it will depend on the status provided to such languages in future reforms of the institutional language regime (Urrutia, 2005).

3.3. The Global Trends to 2035 report

In 2019, the European Parliament's Research Service published a geo-politics and international power report entitled "Global Trends to 2035". This document, written by Oxford Analytica, provides a general overview and scope. It highlights eight trends, among which we find, as trend number three, the "Industrial and technological revolution". According to the report, among the technologies that would bring about an industrial and technological revolution is "Artificial intelligence and automation". The report states that "*One of the largest problems Europe will face in the next two decades is that most of the largest tech providers in the world are based in the United States and China, and their dominance in the sector will be consolidated by the shift to AI*". It raises a question that is as important as it is difficult: how are to be handled the big multinationals that control the field of artificial intelligence? And, consequently, in the face of this oligopoly, what policy should guide

13 https://europa.eu/european-union/sites/europaeu/files/docs/body/treaty_establishing_a_constitution_for_europe_en.pdf

public administration? In line with what the report suggests, we believe that, on the scale that concerns us, of course, one of the principles is "*to guarantee the public nature of data and resources, as well as to encourage the work of local companies that can facilitate technological sovereignty*".

3.4. A new broader multilingualism policy?

We see that both concepts, practical multilingualism and legal multilingualism in Europe are relative, and that in practice, the reason to offer (or not) official information in one of the languages is highly dependent on its cost of translation. However, now, in the context of a constantly evolving Europe, and in the context of the current technological revolution in machine translation and language technology, we have a chance for a new broader policy of multilingualism. It may be defined to capitalise on Europe's linguistic diversity and to encourage the cooperation among local companies that can facilitate technological sovereignty. This has been described as controversial at times and, in fact, the idea has been rejected in some other areas where it has been proposed. It has also been described as necessary for the recognition of different groups or as an advantage for the country in presenting itself to outsiders.

All in all, however, there is no doubt that the recent significant technological advances can be used in a positive way to promote the use of under-resourced languages in Europe. In the current technological scenario, the cost of recognising these languages as official could be affordable, especially since the areas in which each language would be officially used can be dynamically adapted according to demand.

4. Conclusions

Certainly, diversity is socially beneficial. Language diversity too. Europe should succeed in managing its heterogeneity, it should set a milestone in the integration of its citizens, becoming a global benchmark. Then, why not grant the official status to more languages at European level? Is it expensive? In 2021, the costs associated with it are no longer an excuse.

Yes, creating sterile translations is a waste of money. Creating translated texts that no one will ever read is a waste of money. Let us consider that first: what should we translate? Translation supply can be adjusted to demand, to the area and to the depth of information required. Translations can be obtained automatically, and decisions about what to translate and for which languages can be obtained automatically. Current technology could allow us to do so.

Let us pay close attention to the latest advances in language-centred Artificial Intelligence. Let us learn from successful experiences not only for highly-resourced language but also for low-resourced languages, as has been done in the preceding lines. And let us, as a result, define the roadmap to new additional EU official languages for 2025.

Acknowledgements
This work was supported by the POCTEFA227/16, FEDER (LINGUATEC: Development of cross-border cooperation and knowledge transfer in language technologies), and by LIBER Strategy (European Language Equality -ELE- Project).

References

Aldabe, Itziar, Josu Aztiria, Francho Beltrán, Myriam Bras, Klara Ceberio, Itziar Cortes, Jean-Baptiste Coyos, Benaset Dazeas, Louise Esher, Gorka Labaka, Igor Leturia, Kepa Sarasola, Aure Séguier, Jean Sibille (2019) LINGUATEC: Desarrollo de recursos lingüísticos para avanzar en la digitalización de las lenguas de los Pirineos. Procesamiento del Lenguaje Natural, (forthcominng) ISSN 1989-7553

Agirrezabal, Lore (2010). *The basque experience : some keys to language and identity recovery*. Eskoriatza, Gipuzkoa: Garabide Elkartea. ISBN: 978-84-613-6835-8 .

Alegria, Iñaki; Sarasola, Kepa (2017). *Language technology for language communities: An overview based on our experience*. In Communities in Control: Learning tools and strategies for multilingual endangered language communities (CinC 2017), October 19-21, Alcanena, Portugal, p 91-97.

Cortes, Itziar, Igor Leturia, Iñaki Alegria, Aitzol Astigarraga, Kepa Sarasola eta Manex Garaio (2018) Massively multilingual accessible audioguides via cell phones EAMT 2018 (Project/Product Track)

EPRS, European Parliamentary Research Service (2019). *Global Trends to 2035. Geo-politics and international power.* https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_STU(2017)603263

Fan, Angela; Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, MandeepBaines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary. *Beyond English-centric multilingual machine translation.* Journal of Machine Learning Research, 22(107):1–48, 2021.

Instituto Cervantes (2015). *Presentación del «Plan de Impulso de las Tecnologías del Lenguaje»*. OCLC 1164865977.

Mineco (2015). *Plan for the Advancement of Language Technology*. Spanish Government.

Urrutia Libarona, I. (2015). Régimen jurídico de las lenguas y reconocimiento de la diversidad lingüística en el Tratado por el que se establece una Constitución para Europa. *Revista de Llengua i Dret*, 42: 231-273.