LANGUAGE TECHNOLOGIES FOR A MULTILINGUAL PUBLIC ADMINISTRATION IN SPAIN

Iria de-Dios-Flores, José Ramom Pichel Campos, Adina Ioana Vladu, Pablo Gamallo Otero*

Abstract

Interactions between citizens and the public administration are increasingly taking place by electronic means, often referred to as e-government. In Spain, these interactions mostly have to be monolingual, in Spanish, in the case of the central administration, but may be bilingual or even multilingual in autonomous communities that have their own official language. In this paper, we intend to show how state-of-the-art oral and written linguistic technologies for Spain's co-official languages could allow speakers of these languages to use them in many of their administrative relations with any Spanish public agency, thereby facilitating the conversion of Spain's mostly monolingual administration into a multilingual one, enhancing digital language equality and guaranteeing the linguistic rights of speakers of minoritised languages. We will present an overview of the most promising language technologies in terms of their relevance from the point of view of multilingual communication between citizens and the administration. We will also review the existing technologies for Spain's co-official languages, and present some ideas on how these could be integrated towards the multilingual transformation of Spanish public administrations without neglecting some of the attendant ethical and legal issues. The present work is intended to serve as an introductory and accessible overview for legislators, administrators, or any other person interested in the potential of language technologies to assist in developing a multilingual public administration.

Keywords: language technologies; public administration; digital language equality; multilingualism.

TECNOLOGIES DEL LLENGUATGE PER A UNA ADMINISTRACIÓ PÚBLICA MULTILINGÜE A ESPANYA

Resum

Les interaccions entre la ciutadania i l'administració pública es produeixen cada cop amb més freqüència per via electrònica, que sovint s'anomena administració electrònica. A Espanya, moltes d'aquestes interaccions han de ser monolingües, en castellà, en el cas de l'administració central, però poden ser bilingües o fins i tot multilingües a les comunitats autònomes amb llengua oficial pròpia. En aquest article, volem mostrar com les últimes tecnologies del llenguatge oral i escrit per a les llengües cooficials d'Espanya permetrien que els parlants d'aquestes llengües les fessin servir en gran part de les seves relacions administratives amb qualsevol organisme públic espanyol, cosa que facilitaria la transformació de l'administració majoritàriament monolingüe d'Espanya en una de multilingüe i així fomentaria la igualtat lingüística digital i garantiria els drets lingüístics dels parlants de les llengües minoritzades. Presentarem un panorama general de les tecnologies del llenguatge més prometedores per la seva importància des de la perspectiva de la comunicació multilingüe entre la ciutadania i l'administració. També analitzarem les tecnologies existents per a les llengües cooficials d'Espanya i presentarem algunes idees sobre com es podrien integrar per avançar cap a la transformació multilingüe de les administracions públiques espanyoles sense oblidar algunes de les qüestions ètiques i jurídiques dels treballadors. Aquest article té l'objectiu de servir com una descripció introductòria i accessible per a legisladors, administradors o qualsevol altra persona interessada en el potencial de les tecnologies del llenguatge per ajudar a desenvolupar una administració pública multilingüe.

Paraules clau: tecnologies del llenguatge; administració pública; igualtat lingüística digital; multilingüisme.

* Iria de-Dios-Flores, Universidade de Santiago de Compostela. iria.dedios@usc.gal. 0000-0002-5941-1707.

José Ramom Pichel Campos, Universidade de Santiago de Compostela. jramon.pichel@usc.gal. 0000-0001-5172-6803.

Adina Ioana Vladu, Universidade de Santiago de Compostela. adina.vladu@usc.gal.

Pablo Gamallo Otero, Universidade de Santiago de Compostela. pablo.gamallo@usc.gal. 0000-0002-5819-2469.

Article received: 14.01.2023. Blind reviews: 22.02.2023 and 06.03.2023. Final version accepted: 03.05.2023.

Recommended citation: De-Dios-Flores, Iria, Pichel Campos, José Ramon, Vladu, Adina Ioana, y Gamallo Otero, Pablo. (2023). Language technologies for a multilingual public administration in Spain. *Revista de Llengua i Dret, Journal of Language and Law*, *79*, 78-97. https://doi.org/10.58992/rld.i79.2023.3943

Contents

1 Introduction

Multilingualism and linguistic diversity are values of undoubted richness and benefits for our societies. Nonetheless, they are also accompanied by great challenges, even more so in the digital age. This is because whether or not languages, and particularly minorized languages, are made available and supported in digital spaces has a great impact on their prestige (a decisive factor in the normalization of the language) and the linguistic rights of citizens, social (in)equality, and the digital divide. The present work focuses on how state-of-the-art of language technologies can facilitate multilingual communication between citizens and public administrations in Spain. Spain is a multilingual state where several co-official languages besides Castilian are spoken in a number of bilingual autonomous communities (Galician in Galicia, Basque in the Basque Country and Navarre, Catalan in the Valencian Community and in the Balearic Islands), while Catalonia is trilingual (Catalan and Aranese). Altogether, the communities where two or more co-official languages are spoken represent 41.34% of the total population, and this is without considering communities whose languages have not yet been made official, such as Asturias (Asturian-Leonese and Galician), Castilla y León (Asturian-Leonese and Galician), Extremadura (Extremaduran and Galician-Portuguese), Aragón (Aragonese and Catalan) or Ceuta and Melilla (variants of Arabic).

Interactions between citizens and the public administration are increasingly taking place by electronic means, often referred to as e-government, allowing citizens to initiate a procedure telematically, from any place, at any time (Damascene Twizeyimana & Andersson, 2019; Gómez-Pomar Rodríguez & López Aranda, 2009). In the case of Spain's central administration, these interactions mostly have to be monolingual in Castilian, but may be bilingual or multilingual in autonomous communities that have their own official languages. In this article, we intend to describe how state-of-the-art oral and written linguistic technologies for all these languages could allow the speakers of minoritised languages to use them in many of their administrative relations with any Spanish public agency. This would facilitate the conversion of a mostly monolingual Spanish administration into a multilingual one, enhancing digital language equality and ensuring the linguistic rights of speakers of minoritised languages.

The reasons behind the requirement to carry out a large proportion of formalities with the central government through Spanish-only electronic channels require careful examination and lie outside the scope of this article. Lack of resources, a tendency to centralisation of information and interoperability restrictions could be cited as some of the possible main reasons (Bernadí-Gil, 2004, 2008). Nonetheless, a deeper integration of language technologies may have a highly positive impact on meeting the commitments set out, among others, in the European Charter for Regional or Minority Languages (1992), such as the objective set out in Article 7 regarding "the facilitation and/or encouragement of the use of regional or minority languages, in speech and writing, in public and private life", which includes the provision of means to ensure that users of regional or minoritised languages may submit oral or written applications, documents or requests, and receive a reply in those languages (Article 10 on administrative authorities and public services sets out detailed recommendations in this regard). Interestingly, such objectives could be fulfilled, at least in part, through the development of technologies such as machine translation, text analysis and dialogue systems, widely present in today's digital society, culture and economy. Deploying these technologies for less economically powerful languages with fewer speakers is a powerful driving force for the democratisation of the communities that use them, due to their great social and cultural impact. For example, dialogue systems allow us to communicate with machines in our own language; automatic translation increases access to content in different languages, thus facilitating multilingualism; and text-to-speech and speech-to-text systems broaden access to technology for different user profiles. All of this would help to guarantee digital inclusion and language equality at a time when a significant number of European languages are in danger of digital extinction (Gaspari et al., 2021, 2022, inter alia).

The contents of this paper are presented as follows. First, in section 2, we give an overview of the most promising language technologies in terms of their relevance from the point of view of ensuring multilingual communication between citizens and the administration. This introductory section will pave the way for section 3, in which we review the current state of language technologies for Spain's co-official languages, paying particular attention to their use by the public administration. In section 4, we present some ideas on how these technologies could be more efficiently integrated to enhance the multilingual transformation of Spanish public administrations. Finally, in section 5, we review some of the risks, ethical and legal issues that

must be considered before deploying these tools for public use. The ultimate goal of this paper is to serve as an introductory and accessible overview for legislators, administrators, or any other person interested in the potential of language technologies to assist in developing a multilingual public administration.

2 Language technologies: an overview

Language technologies are a set of information technology tools that enable computers to process human language in either speech or text form. In this section, we will provide an overview of some of the main tasks and applications in which language technologies have made more progress in recent years, and taking into consideration their potential uses in multilingual administration. These are: (1) machine translation, (2) speech synthesis, (3) speech recognition, (4) dialogue systems, (5) text generation, and (6) automatic correction of spelling, grammar and style. We will also consider hybrid systems that integrate two or more of these applications. Our main goal is to explain, in simple terms, the main applications of these technologies and how they work, focusing on the potentially most useful elements from the point of view of multilingualism, in terms of how citizens communicate with the public administration and vice versa. It is for this reason, and for the sake of clarity and brevity, that technologies such as information extraction, opinion mining and fact checking are not included in this review. Even when these are crucial language technologies which can be leveraged to enhance the services of the public administration, they are not so straightforwardly part of the tools involved in facilitating citizens' communication with the public administration.

As a research field, language technology is often described as a subfield of artificial intelligence (AI), also referred to as natural language processing (NLP), computational linguistics, or language-centric AI, with each term having its own particular nuances. Technologies such as speech recognition, machine translation and chatbots are very much present in every citizen's daily life. These tools have improved impressively in recent years, particularly due to the explosion of deep learning methods. Although language technology research can be traced back to at least the 1950s, it is only during the last couple of decades that the field has taken a significant step forward, with rule-based systems being gradually displaced by data-based systems. Currently, most language technology tools include at least some elements based on deep learning techniques. Without entering into very technical details, deep learning enables "computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction" (Cun et al., 2015:436). In other words, deep learning methods are a family of algorithms that, having been exposed to immense amounts of raw data, are capable of learning complex implicit features and representations of the input, be this text or audio in the case of human language (Brown et al., 2020; Devlin et al., 2019; Radford et al., 2022). Artificial multi-layer neural networks, including transformer models (Vaswani et al., 2017), are a subset of such algorithms. Very commonly used in language processing research, these algorithms are responsible for the most advanced implementations of the language technology tasks described in the following subsections (an overview of language processing algorithms can be found in Goldberg, 2017; Pilehvar & Camacho-Collados, 2020; or Tunstall et al., 2022). It is important to note that these state-of-the-art systems require large amounts of data to be trained, posing a challenge for low-resource languages that lack the necessary high-quality textual or speech resources.

2.1 Machine translation

Machine translation is perhaps one of the most widely used language technologies and one of the technological keys to multilingualism. According to statistics published by Google,[1] the company's translation services are used by around 500 million people, translating more than 100 billion words a day. The power of machine translation is indisputable, allowing users around the world to access content and communicate in languages other than their own in a fast and cost-effective way. Its benefits in terms of improved accessibility are beyond question and, in the public sector, it has the potential of making information and services available to a wider audience in multiple languages, and is particularly useful for reaching populations that may not speak the dominant language of a region. The range of applications that integrate machine translation services is continuously growing (e.g., from web browsers to mobile phones) and is expected to continue to expand hand in hand with new developments, integrating other language technologies such as speech recognition and

1 These statistics can be found at https://blog.google/products/translate/ten-years-of-google-translate/

synthesis in hybrid systems such as speech-to-speech translation. Nonetheless, it must be noted that the vast majority of the world's languages (approximately 6,500) are not sufficiently supported by machine translation tools. In the context of the present work, it is particularly interesting to distinguish between domain-general and domain-specific machine translation, where the former is implemented in general-purpose systems and the latter is implemented in cases where a specific linguistic domain of interest has certain peculiarities (e.g., legal, academia, administration, sports, etc.) (Pecina et al., 2015).

From the beginnings of NLP, different strategies have been used to carry out machine translation, which can be divided into two main phases: the initial phase, based on the implementation of linguistic knowledge (lexical, grammatical and even semantic) between a specific language pair, known as ruled-based machine translation (Shiwen & Xiaojing, 2014); and a second phase, based on obtaining huge datasets of quality linguistic translations which are then used to feed machine translation systems, whether statistical (Lopez, 2008; Koehn, 2009) or neural (Koehn, 2020). Large volumes of translations are not necessary for the development of rule-based machine translation systems. Rather, the key issue is the implementation of linguistic knowledge in the language pair for which the machine translator is designed, in the form of equivalence rules. For statistical and neural machine translation systems, however, it is essential to have large volumes of human quality parallel translations between the source and target language (or languages) to be translated. For this reason, although neural machine translation is the state of the art in machine translation for many language pairs (e.g., English-German, English-Spanish or English-French), rule-based machine translation systems are still deployed in certain contexts, where the availability of parallel training data – a prerequisite for statistical and NMT approaches – is limited, or where the need for precise, domain-specific translations is particularly high. Very recently, attempts are being made to take advantage of multilingual models for zero-shot translation (Liu et al., 2020; Bapna et al., 2022), that is, translating between language pairs that a neural machine translation system has never been trained on (Gu et al., 2019). However, the results, although acceptable, are far from the quality achieved with good and large parallel corpora.

2.2 Speech synthesis

Speech synthesis can be defined as the artificial production of human speech, i.e., the automatic generation of speech by machines or computers. Speech may be generated from text or other forms of data, such as phonetic transcriptions or lip movements. In the case of speech generation from text, the most widely used approach, speech synthesis is also known as text-to-speech or text-to-speech conversion. A text-to-speech system takes an arbitrary input text and outputs a speech signal that must meet two fundamental requirements: intelligibility and naturalness. Synthesised speech must therefore be easy to understand as well as similar to that produced by a human being.

Normally, the process of speech synthesis consists of two main stages: a front end and a back end. In the first stage, the input text is transformed into a set of linguistic features that usually include information about phonemes, syllables, words and expressions. In the second stage, the input linguistic features provided by the front end are transformed into the corresponding speech signal.

The advancement of speech synthesis technologies from the previous formant-based parametric synthesis, waveform concatenation, to the use of statistical parametric speech synthesis, greatly enhanced the intelligibility and naturalness of synthesised speech (Agerri et al., 2021). In 2016, DeepMind introduced WaveNet, marking a first for the use of deep learning in speech synthesis. Since then, various deep learning architectures have been developed for speech synthesis, resulting in significant improvement in the quality and naturalness of synthesised voices. The next step comes from end-to-end systems capable of generating speech signals directly from the input text, thus sidestepping the front-end stage (whereby the text is transformed into linguistic features). These advances have also paved the way for the creation of new voice synthesis applications using deep learning techniques (Kumar et al., 2022).

The quality of the artificial voices produced by these systems can be excellent; however, proper training of the voices requires a significant amount of high-quality recorded data. Efforts are currently underway to make these systems more accessible for low-resource languages through methods such as improved data efficiency, transfer learning (Byambadorj et al., 2021) and the training of multilingual models (Wu et al., 2022).

Text-to-speech technology has various practical applications in everyday life. One popular use is the read-aloud tool for web pages, text messages, emails, e-books and other types of digital texts. This technology can also facilitate real-time spoken commentary for live events, thus ensuring accessibility for individuals with reading disorders or visual impairments, as well as voice-controlled assistants, which utilise this technology to convert written text into speech for responding to users' commands and requests, providing information and other tasks. Other applications of text-to-speech technology include speech-based user interfaces, automated customer service systems and assistive technologies for individuals with different types of impairments.

2.3 Speech recognition

Speech recognition, also known as automatic speech recognition, describes the task of converting speech into a computer- or machine-readable transcript. More specifically, the main objective of an automatic speech recognition system is to transform an input speech signal into a sequence of words or characters. Since the most common modality is to convert spoken words into their textual transcription, such systems are also referred to as speech-to-text systems.

In recent years, with the advent of the deep learning era, neural networks have been shown to significantly improve the task of speech recognition. Different architectures such as convolutional neural networks or recurrent neural networks have been applied. More recently, networks known as transformers, first introduced in machine translation and currently widely used in NLP, have become highly successful.

Unsupervised pre-training techniques such as Wav2Vec 2.0 (Baevski et al., 2020) have improved the state of the art, especially in low-data settings (Radford & Kim, 2022). To achieve good performance; however, such techniques generally need a fine-tuning stage, i.e., training on a specific task or dataset. The next step comes from systems such as OpenAI's Whisper, which uses weakly supervised speech recognition to train models where fine-tuning on specific datasets is not needed to achieve high-quality results (Radford & Kim, 2022). Moreover, work on Whisper is focusing on multitask and multilingual models, which may be an opportunity for low-resource languages in all areas, i.e., not only speech, but also text (Conneau et al., 2019).

Automatic speech recognition technology has many practical applications, including subtitling services, speech-to-text translation, voice search, speech-based user interfaces for mobile devices or computer systems, and accessibility tools for individuals with hearing impairments. This technology is also integrated into voice-controlled assistants to enable the conversion of users' spoken commands and requests into written text, which is then used to perform various tasks such as answering questions, playing music or controlling smart home devices.

2.4 Dialogue systems

Dialogue systems, also known as conversational agents or chatbots, are a technology that allows humans to converse with computers in natural language (McTear, 2020). Even though early forms of this technology were only available in text, advances in speech recognition and synthesis have made it possible for users to interact using only their voice. In the construction of conversational or dialogue systems, a distinction is usually made between task-oriented and therefore domain-specific systems (e.g., booking hotels or flights, customer support, etc.) and more general, open-scope systems capable of performing different tasks and carrying on extended conversations on a variety of topics (e.g., Google Assistant, Alexa, etc.). There are many different types of dialogue systems, ranging from simple, rule-based systems capable of handling a limited set of predefined queries to more advanced, machine learning-based systems able to cope with a wide range of inputs and adapt to users' needs (Patlan et al., 2021). Among the latter group, some of the latest generation systems to have emerged recently, such as ChatGPT or LaMDA, appear to be surprisingly capable of holding conversations on any subject in a seemingly smooth and realistic way. Nonetheless, in contrast with more conservative conversational agents based on highly controlled patterns, such as Google Assistant or Alexa, large language model-based systems such as ChatGPT and LaMDA have not yet resulted in real products due to the reputational, ethical and other risks related to the current state of artificial language models, which may produce erroneous, biased, or other harmful content.

On the one hand, we have what is known as social chit-chat, which requires an intense process of language refinement. Due to its informality, this type of response must not only be appropriate, it must also be generated responsibly, in a way that will not negatively affect the reputation of the AI-based service. On the other, the fact of being able to extend conversations across multiple interactions through time, dealing coherently with the history of the interactions as well as the associated information, poses a further scientific challenge (Williams et al., 2013). In addition to the refinement of conversation state tracking techniques, a current trend combines machine learning with knowledge representation (e.g., semantic networks and graphs) to reduce the amount of linguistic data and leverage the existing resources available for other languages using machine translation techniques.

Conversational agents can be used in a wide range of settings to provide information and assist citizens in a variety of tasks such as informing about services or policies, handling queries, scheduling appointments, paying fines or collecting data via surveys or polls.

2.5 Automatic text generation

Natural language generation is the area of NLP that investigates the models, tools, systems and applications that automatically create texts from textual, numerical, graphic or other data. Among the most successful generative language models, GPT-3 (Brown et al., 2020) and T5 (Raffel et al., 2020) stand out. Depending on the type of source data and the goal of the generation, there are usually two main categories (Gatt & Krahmer, 2018). On the one hand, text-to-text systems aim to generate coherent texts from other textual sources. They also often require natural language understanding and include the so-called summarisation of texts, in its extractive and generative facets. The goal here is to merge or summarise related texts or sentences to make them more concise; and to simplify complex texts to, for example, make them more accessible to certain readership profiles. These techniques are also closely related to the spelling and grammar correction of texts, automatic translation, or the generation of automatic paraphrases and questions from texts. On the other hand, data-to-text systems aim to generate understandable texts from non-linguistic data, such as numerical data, graphs, tables, images or other data sources. The generated texts must be adapted to the informational needs and the comprehension capacity of the audience to ensure they adequately convey the information.

Text generation technologies allow generating summaries of texts to be generated by abstracting the meaning; texts of complex documents such as laws to be simplified, making them accessible to a wider audience; text reports to be created from data available in charts; the production of automatic image captions, among many other applications. Crucially, when combined with speech synthesis systems, automatic language generation systems can also be immensely helpful for visually impaired people, as well as helping to overcome the limitations of small display devices (mobiles, tablets, etc.) in which graphical visualisation is not easy. Their ability to adapt information to increase accessibility is a fundamental aspect of automatic language generation technologies.

2.6 Automatic correction of spelling, grammar and style

Automatic correction refers to the use of algorithms that allow computers to verify, correct and evaluate texts automatically, using techniques able to identify errors or deviations including spelling mistakes, grammatical errors, misused words or phrases, tone, fluency, etc. As for the tasks previously described, a variety of techniques are being used for automatic correction in language, depending on the specific goals of the system and the types of resources available. Some common techniques include or combine rule-based systems that use a set of predefined rules or patterns (e.g., grammatical or spelling); dictionary-based systems that use a dictionary or lexicon to identify mistakes or misused words; or deep learning systems that use statistical models trained on large sets of data (Cheng et al., 2022). Linguistic correction and evaluation is not limited to the identification of errors, however; it also allows users to identify appropriate or preferred linguistic structures, content and lexical-semantic density, and the coherence and fluency of a text.

Error correction is a crucial tool that can be integrated into most systems or applications used to handle written language (browsers, email, office software, online forms, etc.) or systems that use natural language (machine translation services, text-to-speech software, chatbots, etc.) to improve the accuracy and clarity of

written texts. One use case of automatic correction of social relevance is a correction/suggestion system that encourages inclusive language.

3 Language technologies in Spain's co-official languages

The different language technologies described in section 2 are currently available with varying degrees of development for the three main co-official languages in Spain: Galician, Basque and Catalan. Detailed reports on the degree of technological progress for each of these languages can be found in Ramírez-Sánchez and García-Mateo (2022) for Galician, Sarasola et al. (2022) for Basque, and Melero et al. (2022a) for Catalan, all created under the auspices of a large-scale project in preparation for the European Language Equality Programme, geared towards the development of strategic research, an agenda and a roadmap for achieving full digital language equality in Europe by 2030.[2] In this section, we will focus on describing some of the main language technology developments for Spain's co-official languages, paying particular attention to successful cases in which these language tools are currently being used by the public administrations – both central government and autonomous communities – to facilitate a relationship of equality for all languages. In this regard, machine translation tools become a key factor in facilitating multilingual communication.

Given that machine translation is the cornerstone of multilingual language technologies, it is important to begin by mentioning PLATA, a machine translation platform implemented in the Spanish central administration. PLATA is designed to be integrated with the web portals of the various Spanish public administrations, allowing these portals to be translated via different web services into Spain's co-official languages (Galician, Basque, and Catalan in the variants of Catalonia and Valencia), as well as into English, French and Portuguese. The platform is based on the rule-based and statistical open-source translation engines Apertium (Alegría-Loinaz et al., 2006; Forcada et al., 2011) and Moses (Koehn, 2007), respectively, depending on the language pair to be translated, and has recently included the implementation of neural models for some language pairs. Specifically, Apertium is used to translate between Spanish and closely related languages (Galician, Catalan/Valencian, French and Portuguese); while distant languages (Basque and English) are translated using language models built initially with the statistical translation system Moses, and currently with the neural translation system Fairseq. PLATA has a post-editing system for words that should not be translated or require a special translation. It also collects translation suggestions from users which are then fed back to the machine translation system to improve its quality and accuracy. The main portals using this service are the Secretary of State for Public Administration, the General Comptroller of the State Administration, MUFACE, the Ministry of Justice, the Ministry of Industry, Trade and Tourism, Ministry of Defence and the Spanish Data Protection Agency.

3.1 Galician

Galician is a language spoken in Galicia, an autonomous community in northwest Spain; it is a Romance language that is closely related to Portuguese and spoken by about 2.4 million people. Various NLP technologies developed for the Galician language by Galician NLP research groups, Big Tech companies (e.g., Microsoft and Google), and small and medium-sized enterprises (e.g., imaxin|software) have been implemented or used by public administrations. For example, the machine translation service Gaio, offered by the Xunta de Galicia, allows users to translate between Galician, Spanish and Portuguese; the system is based on Opentrad/Apertium, an open-source machine translation engine. Customised machine translation systems based on improvements of the same technology are in use in a number of provincial councils and city councils, allowing citizens to access the contents of the Official Gazette of the province (e.g., BOP of Pontevedra) in both Galician and Spanish. With respect to natural language generation, another important example of linguistic technology is the automatic generation of Galician weather forecasts using SimpleNLG, a system which has been adapted to Galician (Cascallar-Fuentes et al., 2018).

Free software developments are available in terms of oral interaction technologies (speech synthesis and recognition), such as Cotovia (Rodriguez Banga et al., 2012), that facilitate the integration of speech synthesis and recognition systems in the Galician public administration. Thus, the website of the Galician government

2 A similar report for Spanish was written by Melero et al. (2022b).

(Xunta de Galicia) allows users to listen to its contents using a text-to-speech system, while the Galician health service (SERGAS) enables interaction in Galician with a speech recognition system both for health care appointments and for clinical professionals (CMDVOZ). Finally, the Galician public administration uses different commercial language technology products and software which include Galician, such as Office and OpenOffice spell-checkers, the spell-checker engine Galgo, the grammar checker Golfiño, and the language style checker Exeria.

Furthermore, a highly ambitious project called Proxecto Nós, promoted by the Xunta de Galicia and led by two research centres attached to the University of Santiago de Compostela (CITIUS and ILG), is currently working towards developing language technologies using AI techniques (de-Dios-Flores et al., 2022). This initiative will facilitate the creation of advanced tools for the Galician public administration by developing all the technologies needed for state-of-the-art speech synthesis and recognition, neural machine translation, dialogue systems, automatic text generation and automatic correction. Worthy of particular note in the Nós Project is the development of a Spanish/English-Galician neural machine translation system. This will be a free, open-source system with the potential to be adapted to the administrative domain (Ortega et al., 2022).[3] With respect to speech technologies, two text-to-speech models have been developed using advanced deep learning architectures based on neural networks; these will be freely available for public use and can be integrated into public administration services. The outputs of the Nós Project can be consulted in its different repositories (GitHub, HuggingFace, and Zenodo).

3.2 Basque

Basque is a language spoken in the Spanish Autonomous Communities of the Basque Country and Navarre, as well as in the regions of Lapurdi, Basse Navarre and Zuberoa in France. Basque, like Galician and Catalan/Valencian, is also present in the central administration's PLATA machine translation system, but it is in the Basque public administration that it is most widely used with different language technology solutions, including machine translation with systems such as Itzuli, which translates Basque into Spanish, French and English. Facilitating spoken interaction between users and the public administration, Itzuli allows the user to talk to the machine translation engine; the translator can then reproduce the translation using speech-to-text and text-to-speech systems. Other systems such as Aditu, developed by the Elhuyar Foundation, have been implemented to automatically add subtitles to videos from the Basque Parliament. In relation to other linguistic technologies related to the improvement of texts in Basque (e.g., spell-checkers), the Basque public administration also has a Linguistic technologies portal which offers access to all these tools.

Finally, in years to come, the various developments produced by the Basque Center for Language Technology (HiTZ), specifically the Gaitu project (similar to Proxecto Nós), will actively promote a multilingual administration in Basque by accelerating the development of AI-based language technology resources.

3.3 Catalan/Valencian/Aranese

Catalan is spoken in several areas, including Catalonia, Valencia, the Balearic Islands, and parts of southern France. It is also spoken by a number of communities in northern Italy and in Andorra. Many different language technologies have been developed for the Catalan language, including NLP tools, machine translation systems and speech recognition software. Besides Catalan and Spanish, Aranese, a variant of Occitan in the Aran Valley of Catalonia, is also official in Catalonia, making Catalonia an officially trilingual community.

All the different machine translation systems for the Catalan language can be accessed from a website promoted by the Catalan public administration. Here users can access an automatic online translator specifically for Catalan-Aranese, as well as other language combinations, including Occitan. The Balearic government has implemented a machine translation system based on Apertium for translation between the Balearic variant of Catalan and Spanish. Finally, for Valencian, the Valencian variant of Catalan, the Generalitat Valenciana has a machine translator called Salt.

3 A demonstration of this tool is available on the following web page: https://demos.citius.usc.es/nos_tradutor

With respect to speech technologies, a variety of speech-to-text and text-to-speech systems have been developed for the Catalan language, including speech recognition, speech synthesis and voice-enabled virtual assistants. These technologies allow users to interact with computers and other devices using spoken Catalan, making it easier and more natural for people to access information and perform tasks.

Worthy of particular note in this regard are recent efforts to obtain free speech resources and software for Catalan. These include Mozilla's Common Voice project; a campaign fostered by the association for the promotion of free software Softcatalà and Projecte Aina, which is generating free spoken resources for Catalan equivalent to 6,458 different voices and 1,200 hours of recorded speech; the contemporary spoken corpus from the Catalan Parliament ParlamentParla, which includes 611 hours of parliamentary speeches; and the neural text-to-speech system Catotron (Külebi et al., 2020).

Lastly, we would like to point out the efforts being made within the Aina project, a far-reaching AI project focused on language technologies, led by the Department of Digital Policies and the Barcelona Supercomputing Centre. Aina is generating high-value open-source linguistic resources which will accelerate the integration of state-of-the-art language technology for Catalan in all Catalan-language public administrations. Notably, Valencia has the Vives project specifically for Valencian Catalan (Valencian).

4 Language technologies to enhance a multilingual administration in Spain

The Spanish central administration is far from being a multilingual administration that allows citizens to express themselves freely in co-official languages, whether orally or in writing, despite the potential for this created by today's intelligent language technologies. Although the review presented in section 3 is far from comprehensive, it nonetheless becomes clear that Spain's co-official languages are endowed with basic language technologies with a level of development sufficient to enhance a multilingual relationship between citizens and governmental bodies if integrated across the board. Unfortunately, at present the integration of these technologies is both scarce and scattered. Rather than offering a comprehensive proposal or structured action plan, which would fall outside of the scope of this paper, in this section we present a series of reflections and ideas for improvement that may prove useful to administrators in promoting a wider implementation of language technologies across public e-government services. Our suggestions for the way in which language technologies could be leveraged to enhance a multilingual administration are structured around two main ideas. On the one hand, it is fundamental to make the most of existing advancements by deeply integrating them within e-government services. On the other, it is crucial to promote the development of state-of-the-art technologies and ad hoc solutions for the potential challenges posed by the integration of these systems. We develop these two ideas in the following two subsections.

4.1 Some ideas for a wider integration of language technologies

Although some language technologies are already being used by different administrations, it is our contention that the existent developments are currently underutilised by the public administrations in Spain. Lack of awareness or proper infrastructure may be some of the reasons that prevent a wider implementation of these tools and, in our view, merely integrating all the existing systems across the board would facilitate significant progress towards making the Spanish administration multilingual. In this respect, machine translation is, without question, the core technology that can facilitate multilingual access to information and communication between citizens and the administration. Despite the existence of cases where the different administrations allow access to content in different languages using this service, the integration of machine translation services remains highly deficient.

One key aspect of multilingual e-government is access to information via government web services. At present, systems in place in the Spanish central administration, such as PLATA (discussed above), use machine translation to provide a multilingual version of all the portals of the central administration and allow many of its contents to be listened to in any co-official language. Yet despite the great value of such a powerful tool, its great potential is not being fully realised. The contents of many web portals are currently only available in Spanish, or only partially available in co-official languages. For instance, it is common to find the contents of the main pages in multiple languages (perhaps because these have been manually translated), while others

containing news or newly created content, for example, remain untranslated. This highlights the need for a deeper implementation of translation systems in web services. Official documents such as official gazettes, guides and reports provide another significant example of this. At the moment, although the electronic version of the Spanish Official Gazette (*Boletín Oficial del Estado*) invites users to select different query languages (including Galician, Basque and Catalan), the user selecting any language other than Spanish will be informed that the contents are unavailable. One might think, with good reason, that the contents of a text with such administrative and legal relevance are too sensitive to be translated automatically, or that any error in this regard may have serious consequences. Nonetheless, machine translation solutions are already successfully implemented in similar contexts by many administrations, such as the official gazettes of some multilingual Autonomous Communities, where human translations are not available due to the cost that this would generate and the need for immediacy in dissemination. Provided the public is duly informed that they are accessing machine translated content that has no legal validity and may contain errors, current machine translation engines have a sufficient degree of development to provide multilingual access to public information.

Similarly, all types of direct communications between citizens and the central government (e.g., inquiries, complaints, etc.) must for the most part be carried out only in Spanish, even when, again, machine translation is capable of facilitating multilingual versions of these procedures. When a citizen contacts the administration in a co-official language, the response could contain the translated version in both directions and the original text from both sides, such that it would be easy to detect any communication errors resulting from the use of a machine translation system.

For the tasks exemplified above, existing open-source machine translation systems are recommended because of their interoperability, the possibility of adding improvements, and their suitability for low-resource languages (Forcada, 2006). Many open-source engines, besides those included in PLATA, were created under the auspices of different public initiatives such as the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL, or Plan for the Promotion of Language Technologies), or the many projects reviewed in section 3. For particularly sensitive procedures, a human review and post-editing system should be put in place to improve the process, guarantee transparency, and facilitate accurate and correct communication. These human-reviewed translations would produce valuable data that could be fed back into the system to improve its quality, and would help generate domain-specific, high-quality, multilingual corpora for all the co-official language combinations.

As should by now be evident, machine translation is not the only linguistic technology with tremendous potential for public administrations whose integration is deficient. Speech synthesis and recognition systems are key in improving accessibility to multilingual content, but not all e-government services include them. The use of speech recognition technologies would make it easier for citizens to complete forms or write messages using their voice in any of the co-official languages. The same applies to listening to text using speech synthesis. The speech processing tools available for the co-official languages reported in section 3 have state-of-the-art quality and are open-source tools created by publicly funded initiatives. These tools could, and should, be integrated into all web services and made readily available to citizens, thereby fostering multilingual communication and accessibility. Crowdsourcing projects such as Mozilla's Common Voice are creating rich and diverse speech databases which can help to improve speech recognition software, just as human-reviewed translations can improve machine translation, as described above.

Given that machine translation and speech recognition and synthesis can be considered to be the minimum technological requirements for a multilingual administration, an interesting first step would be to implement an automatic translation system in the general electronic access point of the Government of Spain (Punto de acceso general electrónico del Gobierno) that would allow many procedures to be carried out in any of the co-official languages. In addition to machine translation, speech recognition and speech synthesis, other language technologies could be implemented in the public administration such as spell-, grammar and inclusive language checkers. All the above-mentioned technologies could also be implemented in sensitive public administration domains, such as health systems, facilitating better health communication between public healthcare professionals and patients in their mother tongue. However, it is important to emphasise that none of these proposals can be put into practice effortlessly. The technical, technological and legal aspects of the widespread implementation of language technologies require a rigorous and long-term action plan, as well as the commitment of all public administrations. Some forms of these are already in place.

4.2 Plans and initiatives to foster language technologies

In recent years, Spain has implemented various plans and initiatives to foster language technologies. One of these initiatives was the already mentioned Plan TL, a public initiative carried out between 2015 and 2020 with the general objective of developing the NLP industry, and machine translation and conversational systems in Spain, in Spanish as well as in the co-official languages. The plan is part of the Spain 2025 Digital Strategy, which seeks to transform the country into a digital hub in Europe.

In this context, the initial goals of the Plan TL were to promote the role of the public administration as a driver of the language industry, with the creation of common platforms for NLP and machine translation, and the development of resources aimed at reusing public information, including the creation of data resources (linguistic infrastructures) and software resources. Despite these aims, the results of the Plan TL focused heavily on resources for the Spanish language.[4] At the same time, to the best of our knowledge, although a few years have now passed since the completion of this project, no report has measured the impact of these results with respect to language technologies in Spain and their use by the public administration.

In continuation of Plan TL, the Spanish government implemented the Plan Nacional de Tecnologías del Lenguaje 2022–2025, within the framework of the Proyectos Estratégicos para la Recuperación y Transformación Económica (PERTE), a government initiative promoting strategic projects aimed at recovering and transforming the Spanish economy, including language technologies in the NextGenerationEU-funded strategic programme for the language industry, "PERTE Nueva economía de la lengua" (PERTE-NEL). The objective of PERTE-NEL is to promote innovative initiatives that will improve the technological maturity of all the languages spoken in Spain within the framework of the new digital reality. The linguistic infrastructure thus created will enable a more efficient digitisation of the public administration and the internationalisation of companies, as well as make industry more competitive. The aim is to guarantee citizens' access to more and better public services through the use of AI in administration and develop data-driven public policies to improve transparency and provide valuable information, according to the Plan for the Digitalisation of Public Administrations 2021–2025.

To this end, PERTE-NEL includes and takes advantage of language technology projects already underway in all the co-official languages, as mentioned in section 3 (Aina, Gaitu, Nós, Vives). Up to 2025, these projects will work together to develop multimodal corpora, annotated data, neural multilingual language models, translation engines and speech recognisers, among others, all key resources for the advancement of digital transformation. Thus, PERTE-NEL represents a huge opportunity to promote cutting-edge language technologies for all the official languages of Spain. Such tools and resources, adapted to specific domains (e.g., machine translation systems for legal texts, text classifiers for medical texts, etc.), can be implemented by the NLP industry in the different Spanish public administrations. This could turn the Spanish public administration into a multilingual public administration that guarantees linguistic rights to all Spanish citizens in Spain, irrespective of the language they predominantly speak, as promoted by the European Charter for Regional or Minority Languages.[5]

However, all the advances made through nationwide programmes and plans for the development of language technologies in Spain may be of little use if the tools and resources generated are not integrated into the public administration, if the synergies between research centres and the language technology industry are not fully exploited, and if there is no solid plan in place, including funding, to ensure continuity beyond the temporal scope of programmes such as Plan TL and PERTE-NEL.

Finally, it is important to understand that the implementation of language technologies, especially the most advanced ones, entails risks in the form of ethical and legal issues which we will describe in the following section.

4 The results can be found at the following link: https://plantl.mineco.gob.es/tecnologias-lenguaje/actividades/Paginas/actividades-resultados.aspx

5 The text of the Charter can be consulted at: https://www.coe.int/en/web/european-charter-regional-or-minority-languages/text-of-the-charter

5 Risks: ethical, and legal issues

As argued in the previous sections, in terms of citizen accessibility and inclusivity, the benefits of a multilingual administration that integrates recent language technologies are numerous. Nonetheless, such technologies also share a number of risks and challenges that should be considered to avoid possible ethical and legal issues in the process of building a trustworthy e-administration.

The rapid advancement of AI and deep learning technologies has and will continue to significantly transform society in the coming decades. This transformation will have a significant ethical impact, as these technologies are capable of both improving and disrupting human lives. The legal and ethical issues that arise from the use of AI include privacy, bias, discrimination and trustworthiness, among other aspects crucial to the lives of individuals and society as a whole.

The EU's 2021 Guidance Note on the application of ethical principles in AI[6] states that, in order to ensure fundamental human rights as enshrined in the Charter of Fundamental Rights of the European Union[7] and relevant international human rights law, AI systems must be based on ethical principles such as respect for human agency; privacy, personal data protection and data governance; fairness; individual, social, and environmental well-being; transparency; and accountability and oversight.

As became patently obvious in the recent COVID-19 pandemic crisis, public administrations can no longer put off working towards the goal of adapting to technological advances and digitisation, and that goal can be achieved by integrating AI applications (Sobrino-García, 2021). In the Spanish context, the move towards a digital administration is reflected in the Spanish government's adoption of the "Spain 2025 Digital Strategy"[8] and the "Spanish RDI Strategy in Artificial Intelligence",[9] in line with the Digital Agenda for Europe, which seeks to create secure digital spaces and services. However, in Spanish administrative regulation, no definition of either AI algorithms, or of any legal precepts regarding their application in the public sector (Sobrino-García, 2021), has yet been established.

The necessary technological progress for the improvement of public service may come into conflict with legal safeguards and fundamental rights such as equality, privacy or the protection of personal data; or with principles or obligations of administrative action such as transparency (Capdeferro, 2020). Therefore, the implementation of AI applications in the process of digital transformation of public administrations should be preceded by the necessary risk analysis and impact assessment (Valero Torrijos, 2020).

We summarise below some of the ethical issues regarding applications of AI, with particular regard to language technologies applications and tools.

5.1 Human-computer interaction, transparency and trust

In the use phase of language technology tools, transparency and trust are key factors. As technology advances, the quality of machine-generated output can be nearly indistinguishable from that of human output, though lacking the same degree of trustworthiness (Kamocki & Witt, 2022). For example, "human-like" chatbots based on large language models can provide output that is linguistically sound, but erroneous or even dangerous in content; and neural machine translation systems can experience "hallucinations", that is, produce content that is seemingly fluent and coherent in the target language but completely unrelated to the input in the original language (Bender & Gebru, 2021).

Moreover, studies show that humans tend to anthropomorphise "human-like" technologies, which can lead to an overestimation of their capabilities and unsafe use (Weidinger et al., 2021). This can be a crucial issue

6 Further details can be viewed at: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf

7 The full text of the Charter can be found at: https://www.europarl.europa.eu/charter/pdf/text_en.pdf

8 More information about the strategy can be consulted at: https://portal.mineco.gob.es/en-us/digitalizacionIA/es-digital-2025/Pages/es-digital-2025.aspx

9 A Spanish government report on this strategy is available to view at: https://knowledge4policy.ec.europa.eu/sites/default/files/Spanish_RDI_strategy_in_AI.pdf

in the relationship of citizens with the public administration, as unmet expectations and unsafe use can lead to an important loss of trust. For these reasons, the limits of language technologies should be communicated unambiguously (Weidinger et al., 2021), and artificial output should be clearly marked as such (Kamocki & Witt, 2022). This should help users build reasonable expectations of the technologies employed and prevent liability on the part of the public administration that integrates them. Two particularly relevant issues in this context are accuracy and misinformation.

5.1.1 Accuracy

The accuracy of machine-generated outputs is particularly important in the context of their use in public administrations, as inaccuracy may lead to liability. Neural machine translation systems, for example, can produce fluent sentences with few linguistic errors, while at the same time lacking consistency, making it difficult to ensure textual coherence. Errors such as terminology mistakes, inconsistent translation of homonyms and other types of inconsistencies across sentences can also be common. In some cases, it can be difficult to detect accuracy mistakes in the output such as, for instance, omissions in the context of otherwise fluent and grammatically correct translations (Stefaniak, 2020).

For public administration translations in particular, accuracy and consistency are of the utmost importance, as mistakes in such documents may have legal consequences. It is thus important to clearly mark machine translations as such, inform of their possible limitations, and refer users to the document in the original language in case of doubt, as well as including human post-editing where necessary. If the purpose of machine translations is merely informative (e.g., news or announcements), their release to the public with little to no post-editing, accompanied by a disclaimer, is a fast and uncostly solution. In more sensitive cases, however, human revision and careful post-editing of the machine translated content must be compulsory.

Similar accuracy issues have been reported in the case of automatic speech recognition systems. Complex factors such as background noise, multiple people talking, signal disruption, and distance all affect these systems' ability to accurately process and understand human speech. Errors in texts obtained by automatic speech recognition can range from biases of the recognition model, such as confusion between similar-sounding words, to semantic changes which significantly affect meaning.

5.1.2 Misinformation

As mentioned previously in this work, chatbot technology, while allowing for a remarkable level of humanlike interaction, presents potential risks if implemented without proper safeguards. Systems such as ChatGPT, which generate text in an apparently human manner, are language sequence models that lack any ability to accurately reason about the world or check the truth of their statements. This vulnerability leaves them open to the potential for generating automated misinformation on an unprecedented scale (Marcus, 2022). Such a system, if adopted by public administrations without the proper protective measures, could lead to the dissemination of inaccurate information. If chatbot technology is to be used safely and reliably in public administration, proper safeguards must be implemented to ensure the accuracy and validity of the information being disseminated.

5.2 Data privacy

Large deep learning models are typically trained on publicly available data, but this data may include personal information or sensitive information about individuals or groups. This raises concerns about privacy and the protection of personal data, as well as the potential for misuse or abuse of this information.

In the more specific context of language technology tools that interact directly with users, such as chatbots, voice assistants or spell-checkers, special concern should be taken with regard to the data collected, i.e., no more than essential data should be collected, nor should data be collected without the user's explicit consent (Kamocki & Witt, 2022).

To address concerns about privacy and the protection of personal data, during the development phase of the technology, researchers and developers can take steps to anonymise or de-identify the training data (Tsamados et al., 2022), and/or obtain explicit consent from individuals before using their data for training. This can help

to ensure that personal information is not included in the training data, and that no data is used in a way that violates the privacy rights of individuals. In the use phase, informing users of the type of data being collected and obtaining explicit consent is crucial.

5.3 Bias and discrimination

AI algorithms learn by exposure to massive quantities of data. This means that the output of such algorithms is directly dependent on the quality of the data they are trained on. If the input is biased, this can result in biased output. For example, if the training data includes a disproportionate amount of information from a particular perspective or group, a machine learning model may learn to reflect that bias. This can lead, directly or indirectly, to unfair or harmful treatment of individuals or groups (Tsamados et al., 2022).

One way to avoid bias in deep learning models is to ensure that the training data is diverse, balanced and representative of a wide range of perspectives and experiences. This can help models learn to generate responses that are fair and inclusive and avoid embedded discrimination.

6 Conclusions

In view of the abridged overview of the current degree of development and potential of language technologies for Spain's co-official languages, presented in sections 2 and 3, the key takeaway of the present work is that all these advancements could be instrumental in converting Spain's monolingual administration into a multilingual one at the oral and written level. Machine translation, a key tool for multilingualism, has reached levels of development that were unthinkable 10 years ago and is already being implemented by various public administrations, proving its benefits. Hence, as we propose in section 4, language technologies could be introduced across all aspects of the public administration using the available tools, especially considering the open-source nature of the vast majority of the developments mentioned in the different sections of this paper.

One critical consideration is the constant need for improvement and adaptation that these technologies require, from the point of view of state-of-the-art research as well as with respect to their cross-cutting implementation in public services. This requires the elaboration of a rigorous and long-term action plan, as well as the commitment of all public administrations. In section 4 of this work, we have provided a series of reflections and ideas for improvement, advocating for a wider implementation of language technologies in public services. We are aware of the fact that this work falls short of making a detailed proposal or action plan. This fundamental issue falls beyond the scope of the present work, but should not be neglected. Notable nationwide plans, such as Plan TL or PERTE-NEL, are already in place or are in the process of being implemented. Nonetheless, further technical efforts, political commitment and investments to ensure continuity are needed to ensure that all these advances are transferred into wider opportunities for citizens to communicate with the administration in co-official languages other than Spanish.

Last but not least, it is crucial for legislators and administrators to keep in mind that these technologies are neither perfect nor can they totally substitute the role of humans, and it is essential to carefully consider the ethical and legal issues in the deployment of each particular use case.

7 Acknowledgments

The present work was funded by the project "Nós: Galician in the society and economy of artificial intelligence" (Proxecto Nós: O galego na sociedade e economía da intelixencia artificial, 2021-CP080); an agreement between the Xunta de Galicia and the University of Santiago de Compostela; and grant ED431G2019/04, awarded by the Galician Ministry of Education, University and Professional Training and the European Regional Development Fund (ERDF/FEDER programme).

8 References

Agerri, Rodrigo, Agirre, Eneko, Aldabe, Itziar, Aranberri, Nora, Arriola, Jose Maria, Atutxa, Aitziber, Azkune, Gorka, Casillas, Arantza, Estarrona, Ainara, Farwell, Aritz, Iakes, Goenaga, Josu, Goikoetxea, Koldo, Gojenola, Inma, Hernaez, Mikel, Iruskieta, Gorka, Labaka, Lopez de Lacalle, Oier, Navas, Eva, Oronoz, Maite, … Soroa, Aitor. (2021). *European language equality. D1.2: Report on the state of the art in LT and language-centric AI*. European Language Equality

Alegría-Loinaz, Iñaki, Arantzabal-Altuna, Iñaki, Forcada, Mikel L., Gómez-Guinovart, Xavier, Padró-Cirera, Lluís, Pichel-Campos, José Ramom, & Waliño, Josu. (2006). OpenTrad: Traducción automática de código abierto para las lenguas del estado español. *Procesamiento del Lenguaje Natural*, *37*, 357–358.

Baevski, Alexei, Zhou, Henry, Mohamed, Abdelrahman, & Auli, Michael. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, & Hsuan-Tien Lin (Eds.), *Advances in neural information processing systems 33 (NeurIPS 2020)*. Curran Associates

Bapna, Ankur, Caswell, Isaac, Kreutzer, Julia, Firat, Orhan, van Esch, Daan, Siddhant, Aditya, Niu, Mengmeng, Baljekar, Pallavi, Garcia, Xavier, Macherey, Wolfgang, Breiner, Theresa, Axelrod, Vera, Riesa, Jason, Cao, Yuan, Chen, Mia, Macherey, Klaus, Krikun, Maxim, Wang, Pidong, Gutkin, Alexander, … Hughes, Macduff. (2022). *Building machine translation systems for the next thousand languages*. Google Research.

Bender, Emily M., Gebru, Timnit, McMillan-Major, Angelina, & Shmitchell, Shmargaret. (2021). On the dangers of stochastic parrots: can language models be too big? *FAccT '21: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Bernadí-Gil, Xavier. (2004). *La incidencia de internet sobre la distribución de competencias*. Observatorio de la Evolución de las Instituciones, Universitat Pompeu Fabra.

Bernadí-Gil, Xavier. (2008). La cooperación interadministrativa y la interoperabilidad. In Agustí Cerrillo i Martínez (Ed.), *Informe sobre la administración electrónica local* (pp. 283–330). Fundació Carles Pi i Sunyer d'Estudis Autonòmics i Locals.

Brown, Tom, Mann, Benjamin, Ryder, Nick, Subbiah, Melanie, Kaplan, Jared D., Dhariwal, Prafulla, Neelakantan, Arvind, Shyam, Pranav, Sastry, Girish, Askell, Amanda, Agarwal, Sandhini, Herbert-Voss, Ariel, Krueger, Gretchen, Henighan, Tom, Child, Rewon, Ramesh, Aditya, Ziegler, Daniel, Wu, Jeffrey, Winter, Clemens, … Amodei, Dario. (2020). Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, & Hsuan-Tien Lin (Eds.), *Advances in neural information processing systems 33 (NeurIPS 2020)*. Curran Associates.

Byambadorj, Zolzaya, Nishimura, Ryota, Ayush, Altangerel, Ohta, Kengo, & Kitaoka, Norihide. (2021). Multi-speaker TTS system for low-resource language using cross-lingual transfer learning and data augmentation. *2021 Asia-Pacific Signal and Information Processing Association annual summit and conference (APSIPA ASC)* (pp. 849–853). IEEE.

Capdeferro, Oscar. (2020). La inteligencia artificial del sector público: desarrollo y regulación de la actuación administrativa inteligente en la cuarta revolución industrial. *IDP. Revista de Internet, Derecho y Política*, 30. https://doi.org/10.7238/idp.v0i30.3219

Cascallar-Fuentes, Andrea, Ramos-Soto, Alejandro, & Bugarín-Diz, Alberto. (2018). Adapting SimpleNLG to Galician language. In Emiel Krahmer, Albert Gatt, & Martijn Goudbeek (Eds.), *Proceedings of the 11th international conference on natural language generation* (pp. 67–72). Association for Computational Linguistics.

Cheng, Lanzhi, Ben, Peiyun, & Qiao, Yuchen. (2022). Research on automatic error correction method in English writing based on deep neural network. *Computational Intelligence and Neuroscience*, *3*, 1–10. https://doi.org/10.1155/2022/2709255

Conneau, Alexis, Khandelwal, Kartikay, Goyal, Naman, Chaudhary, Vishrav, Wenzek, Guillaume, Guzmán, Francisco, Grave, Edouard, Ott, Myle, Zettlemoyer, Luke, & Stoyanov, Veselin. (2019). Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, & Joel Tetreault (Eds.), *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 8440–8451). Association for Computational Linguistics.

Constitución Española. (1978, December 29). *Boletín Oficial del Estado*, *311*.

Council of Europe. (1992). *European charter for regional or minority languages (ETS No. 148)*.

Damascene Twizeyimana, Jean, & Andersson, Annika. (2019). The public value of e-government – A literature review. *Government Information Quarterly*, *36*(2), 167–178. https://doi.org/10.1016/j.giq.2019.01.001

de-Dios-Flores, Iria, Magariños, Carmen, Vladu, Adina Ioana, Ortega, John E., Pichel, José Ramom, Garcia, Marcos, Gamallo, Pablo, Fernández Rei, Elisa, Bugarín-Diz, Alberto, González Gamali, Manuel, Barro, Senén, & Regueira, Xosé Luis. (2022). The Nós project: Opening routes for the Galician language in the field of language technologies. In Itziar Aldabe, Begoña Altuna, Aritz Farwell, & German Rigau (Eds.), *Proceedings of the workshop towards digital language equality within the 13th language resources and evaluation conference* (pp. 52–61). European Language Resources Association.

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, & Toutanova, Kristina. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, & Thamar Solorio (Eds.), *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, Volume 1* (pp. 4171–4186). Association for Computational Linguistics

Erjavec, Tomaž, Ogrodniczuk, Maciej, Osenova, Petya, Ljubešić, Nikola, Simov, Kiril, Pančur, Andrej, Rudolf, Michał, Kopp, Matyáš, Barkarson, Starkaður, Steingrímsson, Steinþór, Çöltekin, Çağrı, de Does, Jesse, Depuydt, Katrien, Agnoloni, Tommaso, Venturi, Giulia, Pérez, María Calzada, de Macedo, Luciana D., Navarretta, Costanza, Luxardo, Giancarlo, Coole, Matthew, … Fišer, Darja. (2022). The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*, *57,* 415–448. https://doi.org/10.1007/s10579-021-09574-0

Forcada, Mikel L. (2006, May 22-28). *Open source machine translation: an opportunity for minor languages* [Workshop presentation]. Strategies for developing machine translation for minority languages, 5th SALTMIL Workshop on Minority Languages, LREC 2006, Genoa, Italy.

Forcada, Mikel, L., Ginestí-Rosell, Mireia, Nordfalk, Jacob, O'Regan, Jim, Ortiz-Rojas, Sergio, Pérez-Ortiz, Juan Antonio, Sánchez-Martínez, Felipe, Ramírez-Sánchez, Gema, & Tyers, Francis M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, *25*(2), 127–144.

Gaspari, Federico, Way, Andy, Dunne, Jane, Rehm, Georg, Piperidis, Stelios, & Giagkou, Maria. (2021). *European language equality. D1.1 Digital language equality (preliminary definition)*. European Language Equality.

Gaspari, Federico, Grützner-Zahn, Annika, Rehm, Georg, Gallagher, Owen, Giagkou, Maria, Piperidis, Stelios, & Way, Andy. (2022). *European language equality. D1.3 Digital language equality (full specification)*. European Language Equality.

Gatt, Albert, & Krahmer, Emiel. (2018). Survey of the state of the art in natural language generation: Core tasks, applications, and evaluation. *Journal of Artificial Intelligence Research*, *61*, 65–170.

Goldberg, Yaov. (2017). *Neural network methods for natural language processing*. Springer.

Gómez-Pomar Rodríguez, Juan, & López Aranda, Miguel. (Eds). (2009). *Administración electrónica: El modelo español* (2nd edition). Euroeditions.

Gu, Jiatao, Wang, Yong, Cho, Kyunghyun, & Li, Victor O.K. (2019). Improved zero-shot neural machine translation via ignoring spurious correlations. In Anna Korhonen, David Traum, & Lluís Màrquez (Eds.), *Proceedings of the 57th annual meeting of the Association for Computational Linguistics* (pp. 1258–1268). Association for Computational Linguistics.

Kamocki, Paweł, & Witt, Andreas. (2022). Ethical issues in language resources and language technology – tentative taxonomy. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, & Stelios Piperidis (Eds.), *Proceedings of the thirteenth language resources and evaluation conference* (pp. 559–563). European Language Resources Association.

Koehn, Phillip. (2005). Europarl: A parallel corpus for statistical machine translation. *Proceedings of Machine Translation, Summit X*: *Papers* (pp. 79–86).

Koehn, Philipp. (2009). *Statistical machine translation*. Cambridge University Press.

Koehn, Philipp. (2020). *Neural machine translation*. Cambridge University Press.

Koehn, Philipp, Hoang, Hieu, Birch, Alexandra, Callison-Burch, Chris, Federico, Marcello, Bertoldi, Nicola, Cowan, Brooke, Shen, Wade, Moran, Christine, Zens, Richard, Dyer, Chris, Bojar, Ondrej, Constantin, Alexandra, & Herbst, Evan. (2007). Moses: Open source toolkit for statistical machine translation. In Sophia Ananiadou (Ed.), *Proceedings of the 45th annual meeting of the Association for Computational Linguistics companion volume proceedings of the demo and poster sessions* (pp. 177–180). Association for Computational Linguistics.

Külebi, Baybars, Öktem, Alp, Peiró-Lilja, Alex, Pascual, Santiago, & Farrús, Mireia. (2020, October 25–29). *CATOTRON – A neural text-to-speech system in Catalan* [Conference presentation]. Interspeech 2020, Shanghai, China.

Kumar, Yogesh, Koul, Apesha, & Singh, Chamkaur. (2022). A deep learning approaches in text-to-speech system: A systematic review and recent research perspective. *Multimedia Tools and Applications*, *82*, 15171–15197. https://doi.org/10.1007/s11042-022-13943-4

LeCun, Yann, Bengio, Yoshua, & Hinton, Geoffrey. (2015). Deep learning. *Nature*, *521*, 436–444. https://doi.org/10.1038/nature14539

Liu, Yinhan, Gu, Jiatao, Goyal, Naman, Li, Xian, Edunov, Sergey, Ghazvininejad, Marjan, Lewis, Mike, & Zettlemoyer, Luke. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, *8*, 726–742. https://doi.org/10.1162/tacl_a_00343

Lopez, Adam. (2008). Statistical machine translation. *ACM Computing Surveys (CSUR)*, *40*(3), 1–49.

Marcus, Gary. (2022, December 19). *AI platforms like ChatGPT are easy to use but also potentially dangerous*. Scientific American.

McTear, Michael. (2020). *Conversational AI: Dialogue systems, conversational agents, and chatbots*. Morgan & Claypool Publishers.

Melero, Maite, Figueras, Blanca, Rodríguez, Mar, & Villegas, Marta. (2022a). *European language equality. D1.15. Report on the Catalan language*. European Language Equality.

Melero, Maite, Peñarrubia, Pablo, Cabestany, David, Figueras, Blanca, Rodríguez, Mar, & Villegas, Marta. (2022b). *European language equality. D1.32. Report on the Spanish language*. European Language Equality.

Ortega, John E., de-Dios-Flores, Iria, Gamallo, Pablo, & Pichel, José Ramom. (2022). A neural machine translation system for Galician from transliterated Portuguese text. In Miguel Á. Alonso, Margarita Alonso-Ramos, Carlos Gómez Rodríguez, David Vilares Calvo, & Jesús Vilares (Eds.), *SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations* (pp. 92–95). CEUR Workshop Proceedings.

Patlan, Atharv Singh, Tripathi, Shiven, & Korde, Shubham. (2021). A review of dialogue systems: from trained monkeys to stochastic parrots. *arXiv*, arXiv:2111.01414 [cs.CL].

Pecina, Pavel, Toral, Antonio, Papavassiliou, Vassilis, Prokopidis, Prokopis, Tamchyna, Aleš, Way, Andy, & van Genabith, Josef. (2015). Domain adaptation of statistical machine translation with domain-focused web crawling. *Language Resources and Evaluation*, *49*(1), 147–193. https://doi.org/10.1007/s10579-014-9282-3

Pilehvar, Mohammad Taher, & Camacho-Collados, Jose. (2020). *Embeddings in natural language processing: theory and advances in vector representations of meaning*. Springer.

Radford, Alec, Kim, Jong Wook, Xu, Tao, Brockman, Greg, McLeavey, Christine, & Sutskever, Ilya. (2022). Robust speech recognition via large-scale weak supervision. *arXiv*, arXiv:2212.04356 [eess.AS].

Raffel, Colin, Shazeer, Noam, Roberts, Adam, Lee, Katherine, Narang, Sharan, Matena, Michael, Zhou, Yanqi, Li, Wei, & Liu, Peter J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1−67.

Ramírez-Sánchez, José Manuel, & García Mateo, Carmen. (2022). *European language equality. D1.15. Report on the Galician language*. European Language Equality.

Rikters, Matiss. (2018). Impact of corpora quality on neural machine translation. In Kadri Muischneck & Kaili Müürisep (Eds.), *Human language technologies – The Baltic perspective* (pp. 126–133). IOS Press.

Rodríguez Banga, Eduardo, García-Mateo, Carmen, Méndez-Pazó, Francisco, González-González, Manuel, & Magarinos, Carmen. (2012). Cotovía: An open-source TTS for Galician and Spanish. In Doroteo Torre Toledano et al. (Eds.), *Proceedings IberSPEECH 2012: "VII Jornadas en Tecnología del Habla" and "III Iberian SLTech Workshop"* (pp. 308–315). Universidad Autónoma de Madrid.

Sarasola, Kepa, Aldabe, Itziar, Diaz de Ilarraza, Arantza, Estarrona, Ainara, Farwell, Aritz, Hernaez, Inma, & Navas, Eva. (2022). *European language equality. D1.15. Report on the Basque language*. European Language Equality.

Shiwen, Yu, & Xiaojing, Bai. (2014). Rule-based machine translation. In Sin-Wai Chan (Ed.), *Routledge encyclopedia of translation technology* (pp. 224–238). Routledge.

Sobrino-García, Itziar. (2021). Artificial intelligence risks and challenges in the Spanish public administration: An exploratory analysis through expert judgements. *Administrative Sciences*, *11*(3), 102. https://doi.org/10.3390/admsci11030102

Stefaniak, Karolina. (2020). Evaluating the usefulness of neural machine translation for the Polish translators in the European Commission. In André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, & Mikel L. Forcada (Eds.), *Proceedings of the 22nd annual conference of the European Association for Machine Translation* (pp. 263–269). European Association for Machine Translation.

Tsamados, Andreas, Aggarwal, Nikita, Cowls, Josh, Morley, Jessica, Roberts, Huw, Taddeo, Mariarosaria, & Floridi, Luciano. (2022). The ethics of algorithms: key problems and solutions. *AI & Society*, *37*, 215–230. https://doi.org/10.1007/s00146-021-01154-8

Tunstall, Lewis, von Werra, Leandro, & Wolf, Thomas. (2022). *Natural language processing with transformers*. O'Reilly.

Valero Torrijos, Julián. (2020). The legal guarantees of artificial intelligence in administrative activity: Reflections and contributions from the viewpoint of Spanish administrative law and good administration requirements. *European Review of Digital Administration & Law*, *1*(1–2), 55–62.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Łukasz, & Polosukhin, Ilia. (2017). Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30 (NIPS 2017)* (pp. 5998–6008). Curran Associates.

Weidinger, Laura, Mellor, John, Rauh, Maribeth, Griffin, Conor, Uesato, Jonathan, Huang, Po-Sen, Cheng, Myra, Glaese, Mia, Balle, Borja, Kasirzadeh, Atoosa, Kenton, Zac, Brown, Sasha, Hawkins, Will, Stepleton, Tom, Biles, Courtney, Birhane, Abeba, Haas, Julia, Rimell, Laura, Hendricks, … Gabriel, Iason. (2021). Ethical and social risks of harm from language models. *ArXiv*, arXiv:2112.04359 [cs.CL].

Williams, Jason, Raux, Antoine, Ramachandran, Deepak, & Black, Alan. (2013). The dialog state tracking challenge. In Maxine Eskenazi, Michael Strube, Barbara Di Eugenio, & Jason D. Williams (Eds.), *Proceedings of the SIGDIAL 2013 conference* (pp. 404–413). Association for Computational Linguistics.

Wu, Jilong, Polyak, Adam, Taigman, Yaniv, Fong, Jason, Agrawal, Prabhav, & He, Qing. (2022). Multilingual text-to-speech training using cross language voice conversion and self-supervised learning of speech representations. *ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8017–8021). IEEE.