EUROPEAN LANGUAGE EQUALITY

D1.32

Report on the Spanish Language

Authors	Maite Melero, Pablo Peñarrubia, David Cabestany, Blanca C. Figueras, Mar Rodríguez, Marta Villegas

Dissemination level	Public
Date	28-02-2022

About this document

Project	European Language Equality (ELE)
Grant agreement no.	LC-01641480 – 101018166 ELE
Coordinator	Prof. Dr. Andy Way (DCU)
Co-coordinator	Prof. Dr. Georg Rehm (DFKI)
Start date, duration	01-01-2021, 18 months
Deliverable number	D1.32
Deliverable title	Report on the Spanish Language
Type	Report
Number of pages	25
Status and version	Final
Dissemination level	Public
Date of delivery	Contractual: 28-02-2022 – Actual: 28-02-2022
Work package	WP1: European Language Equality – Status Quo in 2020/2021
Task	Task 1.3 Language Technology Support of Europe's Languages in 2020/2021
Authors	Maite Melero, Pablo Peñarrubia, David Cabestany, Blanca C. Figueras, Mar Rodríguez, Marta Villegas
Reviewers	Itziar Aldabe, Victoria Arranz
Editors	Maria Giagkou, Stelios Piperidis, Georg Rehm, Jane Dunne
EC project officers	Susan Fraser, Miklos Druskoczi
Contact	European Language Equality (ELE) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland Prof. Dr. Andy Way – andy.way@adaptcentre.ie European Language Equality (ELE) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany Prof. Dr. Georg Rehm – georg.rehm@dfki.de http://www.european-language-equality.eu

Consortium

1	Dublin City University (Coordinator)	DCU	IE
2	Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Co-coordinator)	DFKI	DE
3	Univerzita Karlova (Charles University)	CUNI	CZ
4	Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, Ton Epikoinonion Kai Tis Gnosis	ILSP	GR
5	Universidad Del Pais Vasco/ Euskal Herriko Unibertsitatea (University of the Basque Country)	UPV/EHU	ES
6	CROSSLANG NV	CRSLNG	BE
7	European Federation of National Institutes for Language	EFNIL	LU
8	Réseau européen pour l'égalité des langues (European Language Equality Network)	ELEN	FR
9	European Civil Society Platform for Multilingualism	ECSPM	DK
10	CLARIN ERIC – Common Language Resources and Technology Infrastructure as a European Research Infrastructure Consortium	CLARIN	NL
11	Universiteit Leiden (University of Leiden)	ULEI	NL
12	Eurescom (European Institute for Research and Strategic Studies in Telecommunications GmbH)	ERSCM	DE
13	Stichting LIBER (Association of European Research Libraries)	LIBER	NL
14	Wikimedia Deutschland (Gesellschaft zur Förderung freien Wissens e. V.)	WMD	DE
15	Tilde SIA	TILDE	LV
16	Evaluations and Language Resources Distribution Agency	ELDA	FR
17	Expert System Iberia SL	EXPSYS	ES
18	HENSOLDT Analytics GmbH	HENS	AT
19	Xcelerator Machine Translations Ltd. (KantanMT)	KNTN	IE
20	PANGEANIC-B. I. Europa SLU	PAN	ES
21	Semantic Web Company GmbH	SWC	AT
22	SIRMA AI EAD (Ontotext)	ONTO	BG
23	SAP SE	SAP	DE
24	Universität Wien (University of Vienna)	UVIE	AT
25	Universiteit Antwerpen (University of Antwerp)	UANTW	BE
26	Institute for Bulgarian Language "Prof. Lyubomir Andreychin"	IBL	BG
27	Sveučilište u Zagrebu Filozofski fakultet (Univ. of Zagreb, Faculty of Hum. and Social Sciences)	FFZG	HR
28	Københavns Universitet (University of Copenhagen)	UCPH	DK
29	Tartu Ulikool (University of Tartu)	UTART	EE
30	Helsingin Yliopisto (University of Helsinki)	UHEL	FI
31	Centre National de la Recherche Scientifique	CNRS	FR
32	Nyelvtudományi Kutatóközpont (Research Institute for Linguistics)	NYTK	HU
33	Stofnun Árna Magnússonar í íslenskum fræðum SAM (Árni Magnússon Inst. for Icelandic Studies)	SAM	IS
34	Fondazione Bruno Kessler	FBK	IT
35	Latvijas Universitātes Matemātikas un Informātikas institūts (Institute of Mathematics and Computer Science, University of Latvia)	IMCS	LV
36	Lietuvių Kalbos Institutas (Institute of the Lithuanian Language)	LKI	LT
37	Luxembourg Institute of Science and Technology	LIST	LU
38	Università ta Malta (University of Malta)	UM	MT
39	Stichting Instituut voor de Nederlandse Taal (Dutch Language Institute)	INT	NL
40	Språkrådet (Language Council of Norway)	LCNOR	NO
41	Instytut Podstaw Informatyki Polskiej Akademii Nauk (Polish Academy of Sciences)	IPIPAN	PL
42	Universidade de Lisboa, Faculdade de Ciências (University of Lisbon, Faculty of Science)	FCULisbon	PT
43	Institutul de Cercetări Pentru Inteligență Artificială (Romanian Academy)	ICIA	RO
44	University of Cyprus, French and European Studies	UCY	CY
45	Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied (Slovak Academy of Sciences)	JULS	SK
46	Institut Jožef Stefan (Jozef Stefan Institute)	JSI	SI
47	Centro Nacional de Supercomputación (Barcelona Supercomputing Center)	BSC	ES
48	Kungliga Tekniska högskolan (Royal Institute of Technology)	KTH	SE
49	Universität Zürich (University of Zurich)	UZH	CH
50	University of Sheffield	USFD	UK
51	Universidad de Vigo (University of Vigo)	UVIGO	ES
52	Bangor University	BNGR	UK

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
AI4EU	AI4EU (EU project, 2019-2021)
ASL	American Sign Language
ASR	Automatic Speech Recognition
AUDIAS	Audio, Data, Intelligence and Speech
BSC	Barcelona Supercomputing Center
CH	Cultural Heritage
CL	Computational Linguistics
CLiC	Center for Language and Computation
DLE	Digital Language Equality
EC	European Commission
ELE	European Language Equality *(this project)*
ELE Programme	European Language Equality Programme *(the long-term, large-scale funding programme specified by the ELE project)*
ELG	European Language Grid (EU project, 2019-2022)
GPU	Graphics Processing Unit
HCI	Human Computer Interaction (see HMI)
HMI	Human Machine Interaction (see HCI)
HPC	High-Performance Computing
IIC	Instituto de Ingeniería del Conocimiento
LiNHD	Laboratorio de innovación en Humanidades Digitales
LLI	Laboratorio de Lingüística Informática
LR	Language Resource/Resources
LSE	Spanish Sign Language
LSF	French Sign Language
LT	Language Technology/Technologies
META-NET	EU Network of Excellence to foster META
ML	Machine Learning

MT	Machine Translation
NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
SME	Small and Medium-sized Enterprise
SR	Speaker Recognition
TALP	Language and Speech Technologies and Applications Center (at UPC)
TTS	Text-to-Speech
UAM	Universidad Autónoma de Madrid
UChile	Universidad de Chile
UNAM	Universidad Nacional Autónoma de México
UNED	Universidad Nacional de Educación a Distancia
UPC	Universitat Politècnica de Catalunya

Abstract

This report has been developed in the framework of the European Language Equality (ELE) project, entrusted with the mission to develop a strategic agenda and roadmap for achieving full digital language equality in Europe by 2030. In recent years, the language technology (LT) field as part of the artificial intelligence (AI) sector, has experienced remarkable progress. The advent of deep learning and neural networks over the past decade, together with the considerable increase in the number and quality of resources for many languages have yielded results never seen before. The wide scope of language applications evidences not only that LT is one of the most relevant technologies for society, but also one of the most important AI areas with a fast growing economic impact.

Advances are occurring at a rapid pace, with new models and techniques appearing every few months, making the old ones obsolete. What persists, however, is the importance of data. Well-regulated open access to language data (text and speech) is recognised as essential for the development of new products, applications and services in any language, also for Spanish, which is the objective of this report. In it, we provide a snapshot of the current situation of LT in Spanish, based on a comprehensive survey of language resources and tools, which have been collected and documented in the European Language Grid (ELG), where further details can be consulted and the resources accessed. While Spanish, being one of the most spoken languages in the world, is not threatened by globalisation in the way other languages are, and is well-supported by large industrial corporations and projects, the gap in number of resources and tools compared to English is still big.

Regarding the situation of LT in Spain, we note that the need for a large coordinated effort focused on this sector as highlighted in the 2012 META-NET report has been positively met by the deployment of the *Plan de Impulso de las Tecnologías del Lenguaje* by the Spanish Government, which started in 2015. This national Plan has already created important resources for Spanish in the form of corpora, models and benchmarking tools. Nonetheless, there are still many untapped silos of public language data (text and speech) due to the reluctance of certain sectors of the Administration to effectively implement the European directives on open data and reuse of public information. With the renewed interest in AI-based technologies and the full implementation of the *Plan de Impulso de las Tecnologías del Lenguaje*, we may expect better-regulated access to public sector data as well as full incorporation of cutting-edge technological solutions using the Spanish language by the Administration, thereby acting as a true driver of demand in the LT sector.

1 Introduction

This study is part of a series that reports on the results of an investigation of the level of support the European languages receive through technology. It is addressed to decision makers at the European and national/regional levels, language communities, journalists, etc. and it seeks to not only to delineate the current state of affairs for each of the European languages covered in this series, but to additionally – and most importantly – to identify the gaps and factors that hinder further development of research and technology. Identifying such weaknesses will lay the grounds for a comprehensive, evidence-based, proposal of required measures for achieving Digital Language Equality in Europe by 2030.

To this end, more than 40 research partners and experts in more than 30 European languages have conducted an enormous and exhaustive data collection procedure that provided a detailed, empirical and dynamic map of technology support for our languages.[1] The re-

[1] The results of this data collection procedure have been integrated into the European Language Grid so that they can be discovered, browsed and further investigated by means of comparative visualisations across languages.

ports have been developed in the framework of the European Language Equality (ELE)[2] project. With a large and all-encompassing consortium consisting of 52 partners covering all European countries, research and industry and all major pan-European initiatives, the ELE project develops a strategic research, innovation and implementation agenda as well as a roadmap for achieving full digital language equality in Europe by 2030.

The present report focuses on the Spanish language and gives a snapshot of the current situation of LT in Spanish, which is a moving target due to the rapid advances of AI-based technologies. In spite of being one of the most spoken languages in the world and benefiting from overall fairly good technology support, the gap in number of resources and tools in Spanish compared to English is still big. It is to be expected that with the stimulus of the *Plan de Impulso de las Tecnologías del Lenguaje (PLanTL)*, in operation since 2015, this gap will progressively decrease, and more data, both from the public and private sectors, will become openly available for research and development of the LT industry in Spanish.

2 The Spanish Language in the Digital Age

2.1 General facts

The Spanish language, also known as *Castilian*, is the most spoken Romance language and, according to Ethnologue, the 4th most spoken language of the world, with 543 million speakers (Eberhard et al., 2021). Spanish is the official language of Spain, where it originated as an evolution of Vulgar Latin, but most Spanish speakers are in the Americas. It is spoken natively by approximately 473 million people across 21 countries. See Table 1 for a demographic distribution per country in percentages.[3] Spanish is commonly used at all levels of education in most of these countries.

In the Americas, Spanish coexists with a multitude of indigenous languages such as Quechua and Aymara in Peru and Bolivia, and Guarani in Paraguay. In Spain, Catalan, Galician and Basque are official in the regions where they are historically spoken, but Spanish is the only official language at national level. Spanish uses the latin script adapted to some phonetic particularities. Its graphic system consists of twenty-seven graphs (including the “ñ”) and five digraphs <gu/gü, qu, rr, ch, ll> with a relation graph-phoneme 27-24. However, it has numerous cases of asymmetry and mismatch between phonemes and graphemes (Villafana, 2015). Those cases of phonetic polygraphy and graphemic polyphony slightly hinder its learning in the regions which have dialects where there is a no frontal-fricative distinction between /z/ and /θ/ (<s>, <z>) (parts of southern Spain, the Canary Islands and the Americas), between /ʎ/ and /j/ (<ll>, <y>) (most of the territories except for small parts of Northern Spain), or between alveolar lateral and alveolar tap /l/, /ɾ/ (<l>,<r>) in the coda. Aside from phonetic variation, other dialectal phenomena affect morphology, such as the use of second person pronouns (notably the use of ustedes in the America vs the use of vosotros in Spain for the plural or vos in Argentina and Uruguay vs tú for the singular).

In phylogenetic terms, Spanish is a language from the Romance genus inside the Indo-European family. As a Romance language, it shares many features with most of the other languages that belong to the same genus. Spanish’s canonical word order in simple sentences is Subject Verb Object (SVO), but it is Ergative-Possessive in Action Nominal Constructions (Koptjevskaja-Tamm, 2013). We can consider Spanish an inflective language for its tendency to use inflective morphemes, with a predominancy of suffixes (Dryer, 2013).

[2] https://european-language-equality.eu

[3] Percentages from (Fernández-Vítores, 2021) and https://www.pewresearch.org/fact-tank/2020/01/06/speaking-the-national-language-at-home-is-less-common-in-some-european-countries/

Country	Native	Officiality
Mexico	96.80%	non-official
Colombia	99.20%	official
Spain	80.00%	official
Argentina	98.10%	non-official
Peru	86.60%	official
Venezuela	97.30%	official
Chile	95.90%	non-official
Guatemala	78.30%	official
Ecuador	95.80%	official
Bolivia	83.00%	official
Cuba	99.80%	official
Dominican Republic	97.60%	official
Honduras	98.70%	official
Paraguay	68.20%	official
Nicaragua	97.10%	official
El Salvador	99.70%	official
Costa Rica	99.30%	official
Uruguay	98.40%	non-official
Panama	91.90%	official
Puerto Rico	99.00%	official
Equatorial Guinea	74.00%	official
United States	13.50%	non-official

Table 1: Spanish-speaking countries (ordered by total population) plus the US

2.2 Spanish in the Digital Sphere

According to the 2021 report of the Instituto Cervantes[4], Spanish is the third most used language on the Internet. It has experienced a growth of 1,511% in the period 2000-2020, compared to a 743% increase for English. This growth is due, above all, to the incorporation of Latin American users. However, its growth potential is still very high due to the limited access still seen in some Spanish-speaking countries. While in Spain Internet penetration is very high (92.6%), the average in the Americas is only 67% (CEPAL, 2021; ONTSI, 2020).

The same report states that in multilingual websites, the use of Spanish ranks fourth (4.1% of multilingual pages offer a Spanish version), ahead of German and French, but far behind Russian (8.4%), which has fewer native speakers. The figures highlight the still limited multilingual dimension of content originally written and produced in Spanish, since only a very small percentage of this content is offered in another language. In comparison, English is the real lingua franca on the Internet, as it is used in 59.4% of multilingual websites. Being able to offer multilingual access to a company's website is a critical asset in e-commerce in order to reach as many markets as possible. The usage of online retail platforms in Spain is still low compared with the European average (67.5% of the population shop online, compared to 91% of the Netherlands), but is steadily growing (an increase of 29% in 2020).[5] In Latin America, the ratio is still lower, on average less than the 50% of the population shopped online in 2020.[6]. As the practice of e-commerce grows in the Spanish-speaking territories, an increase of websites offering multilingual versions of their content is to be expected.

[4] https://cvc.cervantes.es/lengua/espanol_lengua_viva/pdf/espanol_lengua_viva_2021.pdf
[5] https://ecommerce-europe.eu/wp-content/uploads/2021/09/2021-European-E-commerce-Report-LIGHT-VERSION.pdf
[6] https://www.emarketer.com/content/latin-america-ecommerce-forecast-2021

One of the most relevant indicators of the vitality of Spanish on the Internet is its prominence on digital platforms. Currently, Spanish ranks second on the most popular social networks (Facebook, Instagram and Twitter) and streaming platforms (Netflix and Youtube). Youtube, in particular, has now become one of the main dissemination channels for popular culture in Spanish. This platform has made the consumer of audiovisual products in Spanish much less confined to their geographical area of reference, favouring an unprecedented transfer of linguistic phenomena between the different varieties of the language.

In contrast, the Spanish Wikipedia ranks only ninth in number of articles, behind not only some big languages like German and French, but also much smaller ones, like Swedish and Dutch.

With regard to advanced AI applications that use Spanish, most of the technological solutions and products offered by big companies (Google, Amazon, Facebook, Apple and Microsoft) have a Spanish version. Some of them even offer support to dialectal varieties, like Mexican Spanish or peninsular Spanish. However, most of these products work best in their English version and do not offer full functionality in Spanish. For example, most virtual assistants have difficulties in understanding queries in Spanish beyond the most simple ones.[7]

3 What is Language Technology?

Natural language[8] is the most common and versatile way for humans to convey information. We use language, our natural means of communication, to encode, store, transmit, share and process information. Processing language is a non-trivial, intrinsically complex task, as language is subject to multiple interpretations (ambiguity), and its decoding requires knowledge about the context and the world, while in tandem language can elegantly use different representations to denote the same meaning (variation).

The computational processing of human languages has been established as a specialised field known as *Computational Linguistics* (CL), *Natural Language Processing* (NLP) or, more generally, Language Technology (LT). While there are differences in focus and orientation, since CL is more informed by linguistics and NLP by computer science, LT is a more neutral term. In fact, LT is largely multidisciplinary in nature; it combines linguistics, computer science (and notably AI), mathematics and psychology among others. In practice, these communities work closely together, combining methods and approaches inspired by all of them, together making up *language-centric AI*.

Language Technology is the multidisciplinary scientific and technological field that is concerned with studying and developing systems capable of processing, analysing, producing and understanding human languages, whether they are written, spoken or embodied.

With its starting point in the 1950s with Turing´s renowned intelligence machine (Turing, 1950) and Chomsky´s generative grammar (Chomsky, 1957), LT enjoyed its first boost in the 1990s. This period was signalled by intense efforts to create wide-coverage language resources, such as annotated corpora, thesauri, etc. which were manually labelled for various linguistic phenomena and used to elicit machine readable rules which dictated how language can be automatically analysed and/or produced. Gradually, with the evolution and advances in machine learning, rule-based systems have been displaced by data-based ones, i. e. systems that learn implicitly from examples. In the recent decade of 2010s we observed a radical technological change in NLP: the use of multilayer neural networks able to solve

[7] https://elpais.com/retina/2019/10/16/talento/1571218870_674350.html

[8] This section has been provided by the editors. It is an adapted summary of Agerri et al. (2021) and of Sections 1 and 2 of Aldabe et al. (2021).

various sequential labelling problems. The success of this approach lies in the ability of neural networks to learn continuous vector representations of the words (or word embeddings) using vast amounts of unlabelled data and using only some labelled data for fine-tuning.

In recent years, the LT community has been witnessing the emergence of powerful new deep learning techniques and tools that are revolutionising the way in which LT tasks are approached. We are gradually moving from a methodology in which a pipeline of multiple modules was the typical way to implement LT solutions, to architectures based on complex neural networks trained with vast amounts of data, be it text, audio or multimodal. The success in these areas of AI has been possible because of the conjunction of four different research trends: 1) mature deep neural network technology, 2) large amounts of data (and for NLP processing large and diverse multilingual data), 3) increase in high performance computing (HPC) power in the form of GPUs, and 4) application of simple but effective self-learning approaches.

LT is trying to provide solutions for the following main application areas:

- **Text Analysis** which aims at identifying and labelling the linguistic information underlying any text in natural language. This includes the recognition of word, phrase, sentence and section boundaries, recognition of morphological features of words, of syntactic and semantic roles as well as capturing the relations that link text constituents together.
- **Speech processing** aims at allowing humans to communicate with electronic devices through voice. Some of the main areas in Speech Technology are Text to Speech Synthesis, i. e. the generation of speech given a piece of text, Automatic Speech Recognition, i. e. the conversion of speech signal into text, and Speaker Recognition (SR).
- **Machine Translation**, i. e. the automatic translation from one natural language into another.
- **Information Extraction and Information Retrieval** which aim at extracting structured information from unstructured documents, finding appropriate pieces of information in large collections of unstructured material, such as the internet, and providing the documents or text snippets that include the answer to a user's query.
- **Natural Language Generation (NLG)**. NLG is the task of automatically generating texts. Summarisation, i. e. the generation of a summary, the generation of paraphrases, text re-writing, simplification and generation of questions are some example applications of NLG.
- **Human-Computer Interaction** which aims at developing systems that allow the user to converse with computers using natural language (text, speech and non-verbal communication signals, such as gestures and facial expressions). Popular applications within this area are conversational agents (better known as chatbots).

LT is already fused in our everyday lives. As individual users we may be using it without even realising it, when we check our texts for spelling errors, when we use internet search engines or when we call our bank to perform a transaction. It is an important, but often invisible, ingredient of applications that cut across various sectors and domains. To name just very few, in the *health* domain, LT contributes for instance to the automatic recognition and classification of medical terms or to the diagnosis of speech and cognitive disorders. It is more and more integrated in *educational* settings and applications, for instance for educational content mining, for the automatic assessment of free text answers, for providing feedback to learners and teachers, for the evaluation of pronunciation in a foreign language and much more. In the *law/legal* domain, LT proves an indispensable component for several

tasks, from search, classification and codification of huge legal databases to legal question answering and prediction of court decisions.

The wide scope of LT applications evidences not only that LT is one of the most relevant technologies for society, but also one of the most important AI areas with a fast growing economic impact.[9]

4 Language Technology for Spanish

The META-NET White paper on the Spanish Language in the Digital Age (Melero et al., 2012), published almost a decade ago, acknowledged that Spanish was well-supported by large industrial corporations and that a fair amount of resources and state-of-the-art technologies had already been produced and distributed for Spanish, while also pointing out at the huge gap in the number of resources and tools when compared to English. Most significantly, the report stressed the need for a large coordinated effort focused on language technologies as well as open access to large amounts of public data. A few years later, the Spanish government responded to this need by approving an ambitious Plan for the promotion of Language Technologies,[10] which started in 2015. The Plan, which is still underway, has the main objective of creating resources for Spanish, the other languages of Spain, and the impulse of the language industry. One of its most recent outcomes is MarIA, a set of massive language models in Spanish that are being developed by the Text Mining unit at the Barcelona Supercomputing Center (BSC).[11].

The Spanish language extends over a very large geographic area and, consequently, many research centers across this area are devoting efforts to developing resources and tools for Spanish, although Spain still leads these efforts. Being a global language, with hundreds of millions of speakers, the number of unannotated resources (text, and to a lesser extent speech) in Spanish is quite large. However, there is still a lack of high-quality, well-curated, annotated resources, especially available under open-access conditions.

In the upcoming sections, we will review the resources available for Spanish at the moment of writing this report, and the projects and stakeholders that have made them possible, whilst also pointing at possible gaps in the landscape. Note that all resources, tools and applications mentioned in the report have been documented in the European Language Grid (ELG),[12] where further details can be consulted and the resources accessed.

In this section, we will first review existing textual data and tools, and then we will look into available speech data and applications.

4.1 Language Data

Language data, in the form of text and speech corpora, is the most important resource to build language-technology tools. Ideally the corpora should be large, freely available, with open licenses, belonging to a variety of domains, and clean, i. e, ready to be ingested by the machine. Current semi-supervised methods require large unannotated corpora to train and smaller, manually annotated corpora for fine-tuning and evaluation.

[9] In a recent report from 2021, the global LT market was already valued at USD 9.2 billion in 2019 and is anticipated to grow at an annual rate of 18.4% from 2020 to 2028 (https://www.globenewswire.com/news-release/2021/03/22/2196622/0/en/Global-Natural-Language-Processing-Market-to-Grow-at-a-CAGR-of-18-4-from-2020-to-2028.html). A different report from 2021 estimates that amid the COVID-19 crisis, the global market for NLP was at USD 13 billion in the year 2020 and is projected to reach USD 25.7 billion by 2027, growing at an annual rate of 10.3% (https://www.researchandmarkets.com/reports/3502818/natural-language-processing-nlp-global-market).

[10] https://plantl.mineco.gob.es/tecnologias-lenguaje/PTL/Paginas/plan-impulso-tecnologias-lenguaje.aspx

[11] https://temu.bsc.es

[12] https://www.european-language-grid.eu

Task	Number of datasets
Morphosyntactic	100
Named entities	80
Transcription	60
Topic	55
Sentiment/polarity	45
QA/Interactions	35
Anonymisation	15
Disambiguation	15
Temporal expressions	10
Hate speech	10
Summarisation	5
Stance	5

Table 2: Number of annotated corpora in Spanish covering different NLP tasks and annotation types

Monolingual corpora

Current language technologies heavily rely on the use of massive language models trained on very large corpora. For many languages it is difficult to reach the necessary amount of data to build massive monolingual models. This is not the case of Spanish. In fact, there are over 20 textual corpora exceeding 100 million words, with half of them reaching a billion words, such as the Now Corpus,[13] or the BNE Corpus, currently the largest corpus for Spanish, although not yet openly available. Most of these corpora have been built by crawling the web and have been cleaned and tagged using automatic tools. Some of these corpora come from well-edited sources such as newspapers, scientific journals, collections of published books, or Wikipedia. In some cases, available corpora can be consulted but not downloaded, Codicach[14] and CORPES[15] are good examples of this.

Most of the corpora available for Spanish correspond to contemporary language, given that more than 80% contain texts from 2010 to date. In fact 42% have been collected after 2018. Taking into account all corpus modalities, we find that 3/4 are textual and only 1/4 are audio or video. Regarding free availability, half of the total resources are freely accessible for all purposes, around 10% can be accessed for a fee, while the remaining 40% are available for research or non-commercial purposes only. Additionally, it should be noted that only half of the Spanish corpora contain linguistic annotations. Most common annotations are morpho-syntactic tags, like part of speech and lemma.

Annotated corpora are needed to fine-tune pre-trained models for specific downstream tasks (e. g. NER, Sentiment Analysis) or applications (chatbots) as well as for evaluation purposes. Table 2 shows an approximate number[16] of corpora in Spanish tagged for a series of tasks. Morphosyntactic tags, like part of speech and lemma are the most common annotations, as well as named entities and speech transcriptions. Not surprisingly, there are more corpora annotated for traditional NLP tasks, such as Sentiment analysis or Topic detection, than for more recent tasks, such as detection of hate speech or bias.

The Spanish corpora that have been documented in ELG are geographically diverse and come from almost everywhere in the Spanish-speaking territories, although with a dispro-

[13] https://www.corpusdelespanol.org/now/
[14] (Sadowsky, 2006), http://sadowsky.cl/codicach.html
[15] https://www.rae.es/banco-de-datos/corpes-xxi
[16] Numbers are approximate for two reasons: they only reflect what the authors of this report have been able to find, and the information or metadata associated to the resources is not always complete.

Domain	Number of datasets
Law, Politics, Government	70
Health, Medicine, Pharmacy	70
Literature, Philology	50
Linguistics	50
Technology, Computer Science	40
Journalism, Newswires	40
History, Archaeology, Anthropology	35
Cinema, Television, Radio	30
Social Media	30
Science, Innovation	30
Economics, Tourism, Finance	25
Education	20
Biology, Environment, Agriculture	10

Table 3: Number of corpora per domain in Spanish

portionate percentage from Spain (50%), Mexico (12%) and the USA (7%). Colombia, Argentina and Chile contribute with 5% each, and the rest of Americas are marginal (1.9% – 0.5%).

Experimental results show that models fine-tuned to specific domains tend to perform better on that domain than general purpose models. In order to train domain-specific models and tools, domain-specific corpora are required. The number of existing corpora in Spanish for the different domains varies greatly. Thus, while a sizeable amount of corpora on legal and administrative language can be found, other domains are less well represented. Table 3 shows a summary of the Spanish corpora that have been reported and organised by domain.[17]

Bilingual corpora

One of the most popular and widely used language technologies is machine translation. To train machine translation models, bilingual parallel data are needed. Parallel datasets may be built out of multilingual corpora, and Spanish appears in many multilingual corpora. In the resources documented in ELG, Spanish versions of multilingual corpora often appear together with the European official languages and with the three other major languages in Spain (30-35 datasets with Catalan, 20-25 datasets with Basque and 10-15 datasets with Galician). In contrast we have not found relevant parallel corpora with other minority languages present in Spain, such as Asturian, Aragonese, Mirandese and Romani, and very few with indigenous languages of the Americas, such as Nahuatl, Guarani, Quechua or Aymara. There is also a lack of bilingual corpora with languages of migrants (CES, 2019; OIM, 2021; FEM and OIM, 2021).

Spanish Sign Language Resources

Spanish Sign Language (LSE) is the sign language used mainly by Spanish deaf people and people who live or interact with them. Although there are no fully reliable statistics, it is estimated that there are more than 100,000 signers of LSE, 20 to 30% of whom use it as their second language. The main body of lexical origin of LSE comes from French Sign Language

[17] See caveat in note 16

Model	Architecture	Training Corpus	Corpus size (tokens)
RoBERTa-bne	roberta	BNE	135B
GPT-2-bne	gpt-2	BNE	135B
BETO	bert	Spanish Corpora	3B
Bertin	roberta	Span-mC4 (part)	28B
Electricidad	electra	Spanish Corpora	3B

Table 4: Monolingual Language Models in Spanish

(LSF), but in more current times, it is receiving strong lexical influences from the American Sign Language (ASL). At least 3 LSE corpora as well as lexicons and learning resources are documented in ELG.

Language models

Major advances in Natural Language Processing have come from training massive language models that may then be fine-tuned to a variety of downstream tasks. Although Spanish is included in all large multilingual models, such as mBERT and XLM-RoBERTa, large monolingual language models have been proved to outperform multilingual models in many tasks.

In the last couple of years, several large models have been trained for Spanish. Table 4 summarises the most relevant ones, as well as the size of the corpora on which they have been trained. Out of these, RoBERTa-bne and BETO are the most popular BERT-based ones, and GPT2-2-bne is the only generative language model to date.[18]

Language models trained on general domain text need to be adapted to specific domains, such as legal, financial, health, etc. using domain-specific corpora, and fine-tuned to specific tasks such as part-of-speech tagging, named entity recognition and classification, question answering, semantic textual similarity, natural language inference, cyberbullying detection, stance detection, hate speech detection, sentiment analysis, summarisation, etc. For these adaptations annotated and in-domain corpora are needed.

4.2 Language Technologies and Tools

Even though applications based on language models tend to be trained end-to-end, thereby limiting the relevance of typical NLP low-level tasks, such as word tokenisation, segmentation, part-of-speech tagging, parsing, etc., those tasks remain an important process for many applications. There are a number of toolkits and packages that exist, that gather and maintain these tools. Some of the most complete toolkits in Spanish are included in Freeling,[19] SpaCy,[20] UDPipe,[21] LIMA[22] and Connexor.[23] Table 5 lists tasks and their coverage by the aforementioned toolkits.[24] As there are a large number of tools and toolkits in Spanish, we have grouped all those that are smaller or have the tools focused only on more specific tasks in the column *Others*.

As well as these basic language processing tasks, there are numerous tools for common end-user tasks in Spanish, such as spellcheckers, grammar and style-checkers, etc. which

[18] https://github.com/PlanTL-GOB-ES/lm-spanish
[19] https://nlp.lsi.upc.edu/freeling/node/1
[20] https://spacy.io
[21] https://cran.r-project.org/web/packages/udpipe/vignettes/udpipe-annotation.html
[22] https://aymara.github.io/lima/
[23] https://www.connexor.com/nlplib/
[24] See caveat in note 16

Tasks	Freeling	SpaCy	UDPipe	LIMA	Connexor	Others
Tokenisation	✓	✓	✓	✓	✓	45
Sentence segmentation	✓	✓	✓		✓	10
Lemmatisation		✓	✓			45
Stemming						20
Morphologic Analysis	✓	✓	✓	✓	✓	40
Named Entity Recognition	✓	✓	✓	✓	✓	65
Word Sense Disambiguation	✓			✓	✓	15
Semantic Role Labeling	✓			✓		20
PoS-tagging	✓	✓	✓	✓	✓	65
Syntactic Parsing	✓		✓	✓	✓	35
Dependency Parsing	✓	✓	✓	✓		10
Optical Text Recognition					✓	10

Table 5: Available tools for NLP tasks in Spanish

can be integrated in most content management systems. Other tools deal with stylometry, plagiarism, information extraction, sentiment analysis, automatic transcription, etc.

Translation technologies

Spanish is well served by popular machine translation platforms, such as Google Translate,[25] Deepl,[26] or Bing.[27] The quality attained by English – Spanish general purpose automatic translation is very high, and when it is finetuned to the domain it is even higher.[28]

In addition, some open-source initiatives have built downloadable translation models to translate from Spanish to other languages of Spain (Catalan, Basque and Galician). Apertium,[29] a toolbox to create rule-based translation systems, is one of them. Rule-based systems are technologically more primitive than neural ones, but for closely related languages, like Spanish and Catalan, they provide reasonable results. In addition, eTranslation,[30] the MT service provided freely by the European Commission to European Public Administrations and SMEs offers neural-based translation between all European official languages, including Spanish.

All in all, we have documented in the ELG repository around 600 automatic translators that involve Spanish and another language among over 130 world languages. Approximately 70% of them support more than three other languages. Some are free-to-use, but not all are open-source, so that the source code or the underlying corpora cannot be accessed.

Speech data and technologies

Speech recognition and speech synthesis are behind some of the most iconic AI applications, such as virtual assistants and dialogue agents. These applications are essentially trained on audio datasets. The Language and Speech Technologies group of the Universitat Politécnica de Catalunya (TALP-UPC)[31] has built over the years a set of noteworthy speech resources,

[25] https://translate.google.com
[26] https://www.deepl.com/translator
[27] https://www.bing.com
[28] see (Rivera-Trigueros, 2021)
[29] https://www.apertium.org
[30] https://ec.europa.eu/digital-building-blocks/wikis/display/CEFDIGITAL/eTranslation
[31] https://www.talp.upc.edu

Speech Technology	Number of tools
Automatic Speech Recognition	
Speech to speech translation	5
Phonetic Transcription	10
Voice Transcription	10
Subtitling	5
Speech to text	15
Text to speech	
Speech generation	30
Speaker recognition	
Voice biometrics	5
Voice clonning	5

Table 6: Tools for Speech Technologies in Spanish

including TC-Star, a large lexicon of Spanish words containing phonemic transcriptions,[32] although most of these resources are not open access.

Some of the audio corpora available in Spanish have been created for purposes other than building tools for speech recognition or synthesis. Many are used to research in areas such as Sociolinguistics, Phonetics or Spanish as Second Language. These audio records are almost all transcribed and include some information about the speakers and the general topic of the conversation.

There are close to a hundred speech technology tools documented for Spanish, including text to speech (TTS), automatic speech recognition (ASR), and speaker recognition (SR). Table 6 shows a distribution of the documented tools available for each speech-based task.[33]

4.3 Projects, Initiatives, Stakeholders

Institutions, universities, open source initiatives and industries play an important role in developing language technologies for Spanish. They are responsible for creating and distributing language tools and resources.

Regarding national initiatives, the aforementioned *Plan de Impulso de las Tecnologías del Lenguaje*[34] plays a central role promoting the development of language resources in Spain. It is supported by the Secretary of State for Digitalisation and Artificial Intelligence, and through its collaboration with the Text Mining Unit in the Barcelona Supercomputing Center it has produced several relevant assets in the biomedical text mining domain, machine translation, and more recently massive language models, in the MarIA initiative.[35]

Another project, called *Spanish Language and Artificial Intelligence* (LEIA),[36] is also currently underway between the *Real Academia Española de la Lengua*, the institution entrusted with the stability of the Spanish language, and the Big Techs (Microsoft, Amazon, Google, Twitter and Facebook) with the objective of ensuring high quality coverage of the Spanish language by their AI products.

Universities and research centers from the Spanish-speaking world play an important role in the research and generation of resources for Spanish. There are many research groups around the world with a focus on Spanish NLP. In spite of this geographical diversity, most

[32] https://www.talp.upc.edu/page-resources-lists
[33] See caveat in footnote 16
[34] https://plantl.mineco.gob.es/tecnologias-lenguaje/PTL/Paginas/plan-impulso-tecnologias-lenguaje.aspx
[35] https://github.com/PlanTL-GOB-ES/lm-spanish
[36] https://www.rae.es/noticia/que-es-leia

of the larger research groups are based in Spain. The following are some of the more active, among the many that exist:

- the *Audio, Data, Intelligence and Speech* (AUDIAS)[37] (IIC)[38] (LLI)[39] from the Universidad Autónoma de Madrid (UAM);
- the *Center for Language and Computation* (CLiC)[40] of the Universitat de Barcelona (UB);
- the *Language and Speech Technologies and Applications Center* (TALP) of the Universitat Politècnica de Catalunya (UPC);[41]
- the *Laboratorio de innovación en Humanidades Digitales* (LiNHD)[42] of the Universidad Nacional de Educación a Distancia (UNED);
- Other non-Spanish universities that have relevant groups with a focus on Spanish NLP are the Universidad Nacional Autónoma de México (UNAM) and the Universidad de Chile (UChile).

Aside from the big companies in the technology industry there are many small and medium companies developing solutions in Spanish. The top services offered by these companies include customised chatbots, machine translation systems, speech technologies, spell-checkers and specialised tools for linguistic information extraction and management.

Finally, we need to mention the *Spanish Society for Natural Language Processing* (SEPLN).[43] The SEPLN is a non-profit organisation, supported by research groups and NLP industry, created back in 1983 with the purpose to promote teaching, research and development of Spanish NLP. Among the main activities of the SEPLN are the organisation of an annual conference, regularly attended by a number of research groups and companies working in the field; the edition of a biannual journal; and a web with news and information about current issues and a forum for members.

5 Cross-Language Comparison

The LT field[44] as a whole has evidenced remarkable progress during the last years. The advent of deep learning and neural networks over the past decade together with the considerable increase in the number and quality of resources for many languages have yielded results unforeseeable before. However, is this remarkable progress equally evidenced across all languages? To compare the level of technology support across languages, we considered more than 11,500 language technology tools and resources in the catalogue of the European Language Grid platform (as of January 2022).

5.1 Dimensions and Types of Resources

The comparative evaluation was performed on various dimensions:

[37] http://audias.ii.uam.es, the *Instituto de Ingeniería del Conocimiento*
[38] https://www.iic.uam.es/iic/ and the *Laboratorio de Lingüística Informática*
[39] http://www.lllf.uam.es/ESP/
[40] http://clic.ub.edu/en/
[41] https://www.talp.upc.edu
[42] https://linhd.uned.es
[43] http://www.sepln.org/en/sepln
[44] This section has been provided by the editors.

- The current state of technology support, as indicated by the availability of tools and services[45] broadly categorised into a number of core LT application areas:
  - Text processing (e. g., part-of-speech tagging, syntactic parsing)
  - Information extraction and retrieval (e. g., search and information mining)
  - Translation technologies (e. g., machine translation, computer-aided translation)
  - Natural language generation (e. g., text summarisation, simplification)
  - Speech processing (e. g., speech synthesis, speech recognition)
  - Image/video processing (e. g., facial expression recognition)
  - Human-computer interaction (e. g., tools for conversational systems)
- The potential for short- and mid-term development of LT, insofar as this potential can be approximated by the current availability of resources that can be used as training or evaluation data. The availability of data was investigated with regard to a small number of basic types of resources:
  - Text corpora
  - Parallel corpora
  - Multimodal corpora (incl. speech, image, video)
  - Models
  - Lexical resources (incl. dictionaries, wordnets, ontologies etc.)

5.2 Levels of Technology Support

We measured the relative technology support for 87 national, regional and minority European languages with regard to each of the dimensions mentioned above based on their respective coverage in the ELG catalogue. For the types of resources and application areas, the respective percentage of resources that support a specific language over the total number of resources of the same type was calculated, as well as their average. Subsequently each language was assigned to one band per resource type and per application area and to an overall band, on a four-point scale, inspired by the scale used in the META-NET White Paper Series, as follows:

1. **Weak or no support**: the language is present (as content, input or output language) in <3% of the ELG resources of the same type
2. **Fragmentary support**: the language is present in $\geq$3% and $<$10% of the ELG resources of the same type
3. **Moderate support**: the language is present in $\geq$10% and $<$30% of the ELG resources of the same type
4. **Good support**: the language is present in $\geq$30% of the ELG resources of the same type[46]

The overall level of support for a language was calculated based on the average coverage in all dimensions investigated.

[45] Tools tagged as "language independent" without mentioning any specific language are not taken into account. Such tools can certainly be applied to a number of languages, either as readily applicable or following fine-tuning, adaptation, training on language-specific data etc., yet their exact language coverage or readiness is difficult to ascertain.

[46] The thresholds for defining the four bands were informed by an exploratory k-means 4-cluster analysis based on all data per application and resource type, in order to investigate the boundaries of naturally occurring clusters in the data. The boundaries of the clusters (i. e., 3%, 10% and 30%) were then used to define the bands per application area and resource type.

5.3 European Language Grid as Ground Truth

At the time of writing (January 2022), the ELG catalogue comprises more than 11,500 metadata records, encompassing both data and tools/services, covering almost all European languages – both official and regional/minority ones. The ELG platform harvests several major LR/LT repositories[47] and, on top of that, more than 6,000 additional language resources and tools were identified and documented by language informants in the ELE consortium. These records contain multiple levels of metadata granularity as part of their descriptions.

It should be noted that due to the evolving nature of this extensive catalogue and differing approaches taken in documenting records, certain levels of metadata captured are not yet at the level of consistency required to carry out a reliable cross-lingual comparison at a granular level. For example, information captured on corpora size, annotation type, licensing type, size unit type, and so on, still varies across records for many languages, while numerous gaps exist for others. As the ELG catalogue is continuously growing, the comprehensiveness, accuracy and level of detail of the records will naturally improve over time. Moreover, the Digital Language Equality (DLE) metric will allow for dynamic analyses and calculations of digital readiness, based on the much finer granularity of ELG records as they mature.[48]

For the purposes of high-level comparison in this report, the results presented here are based on relative counts of entries in the ELG for the varying types of data resources and tools/services for each language. As such, the positioning of each language into a specific level of technology support is subject to change and it reflects a snapshot of the available resources on January 2022.

That said, we consider the current status of the ELG repository and the higher level findings below adequately representative with regard to the current existence of LT resources for Europe's languages.

5.4 Results and Findings

As discussed above, our analysis takes into account a number of dimensions for data and tools/services. Table 7 reports the detailed results per language per dimension investigated and the classification of each language into an overall level of support.

The best supported language is, as expected, English, the only language that is classified in the *good support* group. French, German and Spanish form a group of languages with *moderate support*. Although they are similar to English in some dimensions (e. g., German in terms of available speech technologies and Spanish in terms of available models), overall they have not yet reached the coverage that English has according to the ELG platform. All other official EU languages are clustered in the *fragmentary support* group, with the exception of Irish and Maltese, which have only *weak or no support*. From the remaining languages, (co-)official at national or regional level in at least one European country and other minority and lesser spoken languages[49], Norwegian and Catalan belong to the group of languages with *fragmentary support*. Basque, Galician, Icelandic and Welsh are borderline cases; while they are grouped in the *fragmentary support* level, they barely pass the threshold from the lowest level. All

[47] At the time of writing, ELG harvests ELRC-SHARE, LINDAT/CLARIAH-CZ, CLARIN.SI, CLARIN-PL and HuggingFace.

[48] Interactive comparison visualisations of the technology support of Europe's languages will be possible on the ELG website using a dedicated dashboard, which dynamically analyses the resources available in the ELG repository, from the middle of 2022 onwards.

[49] In addition to the languages listed in Table 7, ELE also investigated Alsatian, Aragonese, Arberesh, Aromanian, Asturian, Breton, Cimbrian, Continental Southern Italian (Neapolitan), Cornish, Eastern Frisian, Emilian, FrancoProvencal (Arpitan), Friulian, Gallo, Griko, Inari Sami, Karelian, Kashubian, Ladin, Latgalian, Ligurian, Lombard, Lower Sorbian, Lule Sami, Mocheno, Northern Frisian, Northern Sami, Picard, Piedmontese, Pite Sami, Romagnol, Romany, Rusyn, Sardinian, Scottish Gaelic, Sicilian, Skolt Sami, Southern Sami, Tatar, Tornedalian Finnish, Venetian, Võro, Walser, Yiddish.

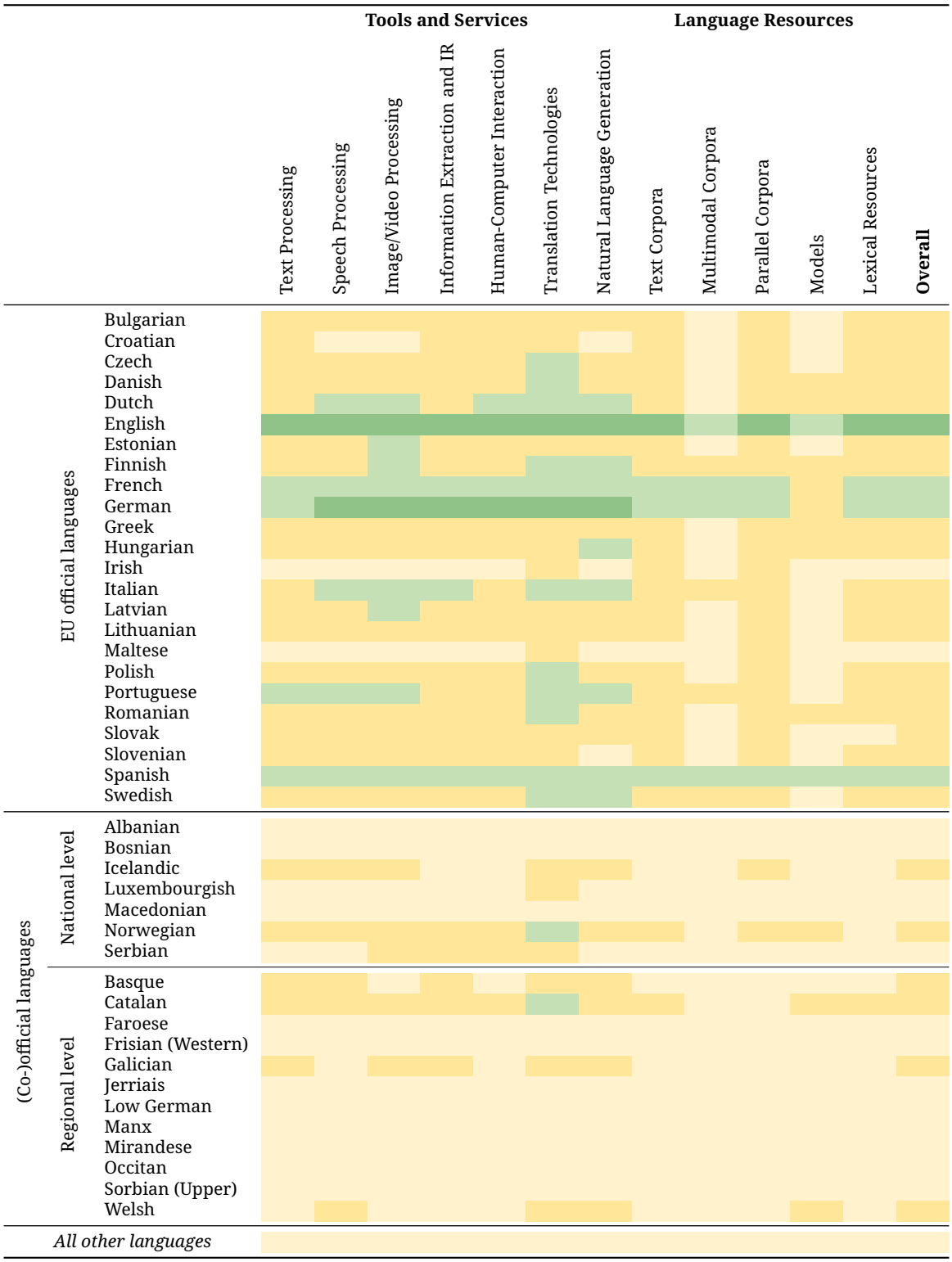

Table 7: State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support)

other languages are supported by technology either weakly or not at all. Figure 1 visualises our findings.

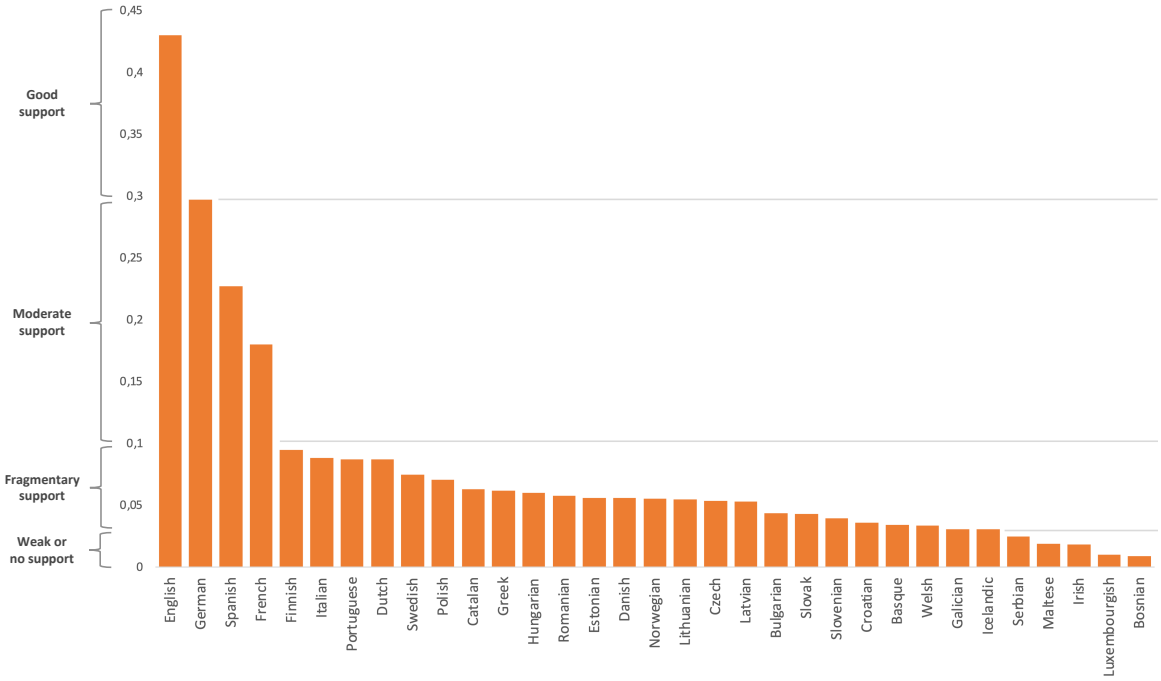

Figure 1: Overall state of technology support for selected European languages (2022)

While a fifth level, *excellent support*, could have been foreseen in addition to the four levels described in Section 5.2, we decided not to consider this level for the grouping of languages. Currently no natural language is optimally supported by technology, i. e., the goal of *Deep Natural Language Understanding* has not been reached yet for any language, not even for English, the best supported language according to our analysis. While recently there have been many breakthroughs in AI, Computer Vision, ML and LT, we are still far from the grand challenge of highly accurate deep language understanding, which is able to seamlessly integrate modalities, situational and linguistic context, general knowledge, meaning, reasoning, emotion, irony, sarcasm, humour, culture, explain itself at request, and be done as required on the fly and at scale. A language can only be considered as excellently supported by technology if and when this goal of Deep Natural language Understanding has been reached.

The results of the present comparative evaluation reflect, in terms of distribution and imbalance, the results of the META-NET White Paper Series (Rehm and Uszkoreit, 2012). The complexities of the analyses clearly differ across 2012 and 2022 studies, and as such, a direct comparison between the two studies can therefore not be made. However, we can instead compare the relative level of progress made for each language in the meantime. It is undebatable that the technology requirements for a language to be considered digitally supported today have changed significantly (e.g. the prevalent use of virtual assistants, chat bots, improved text analytics capabilities, etc.). Yet also the imbalance in distribution across languages still exists.

The results of this analysis are only informative of the relative positioning of languages, but not of the progress achieved within a specific language. The LT field as a whole has significantly progressed in the last ten years and remarkable progress has been achieved for specific languages in terms of quantity, quality and coverage of tools and language resources. Yet, the abysmal distance between the best supported languages and the minimally

supported ones is still evidenced in 2022. It is exactly this distance that needs to be ideally eliminated, if not at least reduced, in order to move towards Digital Language Equality and avert the risks of digital extinction.

6 Summary and Conclusions

In this report we have described the situation of language technologies in Spanish and tried to provide a snapshot of the picture at this point in time. The current moment is an evolving situation where AI-based technologies are heating up, and many languages are trying to catch up with English, in a context of ever-growing globalisation.

Spanish, being one of the most spoken languages in the world, is not threatened by globalisation in the way other languages are. Moreover, due to its specific demographics, its use as a global language in the Internet is expected to grow in the coming years. There is also room for growth in aspects such as multilinguality of Spanish websites, e-commerce, or collaborative initiatives like Wikipedia.

Spanish can also benefit from the global opportunity offered by the latest transformer-based technology which reduces the need for huge amounts of manually annotated data. Thanks to transfer learning and multilingual models, the new models are able to substantially cut down the costs of developing cutting-edge applications.

One big strength of Spanish language technologies is its broad speaker base and wide geographical scope, featuring many research centers who devote their efforts to developing resources and tools for this language, although the bulk of the research is still in Spain. Moreover, many multilingual projects around the world also include Spanish. As noted in previous reports, such as the META-NET White paper on the Spanish Language in the Digital Age (Melero et al., 2012), Spanish is well-supported by large industrial corporations and projects, although the gap in the number of resources and tools compared to English is still big. When it comes to existing resources and applications for Spanish, we have documented a large amount, but, as is the case with other languages, they tend to be scattered throughout many entities and institutions, and not sufficiently accessible. In some cases, even though they have been financed with public money, they are not openly available because they do not have the appropriate licenses. The present survey has also detected deficiencies in diversity of corpora in terms of geographical variation and certain domains and tasks, such as bias detection, conversational systems or anonymisation.

Focusing on the situation of Language Technologies in Spain, the need for a large coordinated effort focused on this sector was already pointed out in the 2012 META-NET report and has been positively met by the deployment of the *Plan de Impulso de las Tecnologías del Lenguaje* by the Spanish Government since 2015. This national Plan has already created important resources for Spanish in the form of corpora, models and benchmarking tools. Nonetheless, there are still many untapped silos of public language data (text and speech) due to the reluctance of certain sectors of the Administration to effectively implement the European directives on open data and reuse of public information. With the renewed interest in AI-based technologies and the full implementation of the *Plan de Impulso de las Tecnologías del Lenguaje*, we may expect better-regulated access to public sector data as well as full incorporation of cutting-edge technological solutions using the Spanish language by the Administration, thereby acting as a true driver of demand in the LT sector.

Acknowledgements

This report has benefited from the insightful comments made by Victoria Arranz and Itziar Aldabe. The authors gratitude goes to them as well as to Maria Giagkou for her diligent monitoring of the edition process.

References

Rodrigo Agerri, Eneko Agirre, Itziar Aldabe, Nora Aranberri, Jose Maria Arriola, Aitziber Atutxa, Gorka Azkune, Arantza Casillas, Ainara Estarrona, Aritz Farwell, Iakes Goenaga, Josu Goikoetxea, Koldo Gojenola, Inma Hernaez, Mikel Iruskieta, Gorka Labaka, Oier Lopez de Lacalle, Eva Navas, Maite Oronoz, Arantxa Otegi, Alicia Pérez, Olatz Perez de Viñaspre, German Rigau, Jon Sanchez, Ibon Saratxaga, and Aitor Soroa. European Language Equality D1.2: Report on the state of the art in Language Technology and Language-centric AI, September 2021. URL https://european-language-equality.eu/wp-content/uploads/2021/10/ELE_Deliverable_D1_2.pdf.

Itziar Aldabe, Georg Rehm, German Rigau, and Andy Way. European Language Equality D3.1: Report on existing strategic documents and projects in LT/AI, November 2021. URL https://european-language-equality.eu/wp-content/uploads/2021/12/ELE___Deliverable_D3_1__revised_.pdf.

CEPAL. Datos y hechos sobre la transformación digital: Informe sobre los principales indicadores de adopción de tecnologías digitales en el marco de la agenda digital para américa latina y el caribe. 2021. URL http://hdl.handle.net/11362/46766.

CES. La inmigración en España: efectos y oportunidades. *Colección Informes*, 02, 2019. URL http://www.ces.es/documents/10180/5209150/Inf0219.pdf.

Noam Chomsky. *Syntactic Structures*. The Hague: Mouton, 1957.

Matthew S. Dryer. Prefixing vs. suffixing in inflectional morphology. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL https://wals.info/chapter/26.

David M Eberhard, Gary F Simons, and Charles D Fennig, editors. *Ethnologue: Languages of the World*. SIL International, Dallas, TX, USA, twenty-fourth edition, 2021.

FEM and OIM. Movimientos migratorios recientes en América del Sur. Informe Anual 2021. 2021. URL https://robuenosaires.iom.int/sites/robuenosaires/files/publicaciones/OIM_Movimientos-Migratorios-FEM-Informe-anual-2021.pdf.

David Fernández-Vítores. El español: una lengua viva. informe 2021. 2021. URL https://doi.org/10.1007/s10579-021-09537-5.

Maria Koptjevskaja-Tamm. Action nominal constructions. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL https://wals.info/chapter/62.

Maite Melero, Toni Badia, and Asunción Moreno. *La lengua española en la era digital – The Spanish Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-30840-6. Available online at http://www.meta-net.eu/whitepapers.

OIM. Tendencias Migratorias en Centroamérica, Norteamérica y el Caribe. 2021. URL https://rosanjose.iom.int/site/sites/default/files/Reportes/sitrep-12_de_oct_2021.pdf.

ONTSI. La sociedad en red. transformación digital en españa. informe anual 2019. 2020. ISSN 1889-9471. URL https://www.ontsi.red.es/sites/ontsi/files/2020-11/InformeAnualLaSociedadEnRed2019Ed2020.pdf.

Georg Rehm and Hans Uszkoreit, editors. *META-NET White Paper Series: Europe's Languages in the Digital Age*, 32 volumes on 31 European languages, Heidelberg etc., 2012. Springer.

Irene Rivera-Trigueros. Machine translation systems and quality assessment: a systematic review. *Language Resources and Evaluation*, 2021. URL https://cvc.cervantes.es/lengua/espanol_lengua_viva/pdf/espanol_lengua_viva_2021.pdf.

Scott Sadowsky. Corpus dinámico del castellano de chile (codicach), 2006. URL http://sadowsky.cl/codicach.html.

Alan Mathison Turing. I.—computing machinery and intelligence. *Mind*, LIX(236):433–460, 10 1950. ISSN 0026-4423. doi: 10.1093/mind/LIX.236.433. URL https://doi.org/10.1093/mind/LIX.236.433.

Jorge Esquivel Villafana. El sistema ortográfico de la RAE (2010): Un estado de la cuestión. *Escritura y Pensamiento*, 18(37):137–152, 2015.