EUROPEAN LANGUAGE EQUALITY

D1.15

Report on the Galician Language

Authors	José Manuel Ramírez Sánchez, Carmen García Mateo
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	D1.15
Deliverable title	Report on the Galician Language
Type	Report
Number of pages	20
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	José Manuel Ramírez Sánchez, Carmen García Mateo
Reviewers	Eva Navas, Sabine Kirchmeier
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
ASR	Automatic Speech Recognition
BERT	Bidirectional Encoder Representations from Transformers
CL	Computational Linguistics
DLE	Digital Language Equality
ELE	European Language Equality *(this project)*
ELG	European Language Grid (EU project, 2019-2022)
GPU	Graphics Processing Unit
GNU/GPL	General Public License
HPC	High-Performance Computing
ICT	Information and communication technology
LT	Language Technology
mBERT	multilingual Bidirectional Encoder Representations from Transformers
META	Multilingual Europe Technology Alliance
META-NET	EU Network of Excellence to foster META
ML	Machine Learning
NLG	Natural Language Generation
NLP	Natural Language Processing
POS	Part-of-Speech
SR	Speaker Recognition
TTS	Text-to-speech

Abstract

This report is part of a series of investigations carried out by the European Language Equality (ELE) project to design a joint agenda and a road-map to achieve full digital language equality in Europe by 2030. The purpose of this report is to show the current state of language technology in terms of resources, services, and community for the Galician language. We find a reduced number of resources, products, and technologies for the Galician language. There are few applications for speech synthesis, speech recognition, spelling correction, grammar, and automatic translation (mostly between Spanish and Portuguese). There are big and high-quality text databases for Galician, but there is a huge gap in multimedia resources. In this scenario, text-based technologies and services are more mature than those based on speech processing. The structure of this document starts with a short introduction, followed by section two, where we discuss some general and formal facts about Galician and its community of speakers in the digital era. Section three is a brief introduction to the field of Language Technology, its main application/research areas, and methodologies. In section four, we present a high-level qualitative description of the resources, projects, initiatives, and stakeholders for Galician. Section five offers a cross-language comparison between Galician and other European languages using metrics developed for the ELE Project. The document ends with a summary and conclusions section.

Extended Abstract

Este informe forma parte dunha serie de estudos levados a cabo polo proxecto europeo European Language Equality (ELE) para deseñar de forma conxunta unha axenda e unha folla de ruta que permitan a plena igualdade lingüística dixital en Europa no ano 2030. O propósito deste informe é mostrar o estado actual das tecnoloxías lingüísticas en canto a recursos, servizos e comunidade de falantes para a lingua galega. Como conclusión xeral pódese dicir que existe un número bastante reducido de recursos, produtos e tecnoloxías para a lingua galega. Hai poucas aplicacións de síntese de voz, recoñecemento de voz, corrección ortográfica e gramatical e tradución automáticas. É ben certo que existen bases de datos de texto en galego de gran dimensión e de gran calidade, pero hai un baleiro importante en canto a recursos multimedia adecuados para desenvolver aplicacións tan importantes no mundo actual como os axentes conversacionais por voz de última xeración. En cambio, as tecnoloxías e os servizos baseados en texto están nunha fase máis madura.

Os resultados do estudio falan dunha situación con moita marxe de mellora para a lingua galega; non só en termos de presenza na Internet, senón tamén no tocante a recursos e soporte dixital. Os datos recompilados mostran unha brecha considerable en comparación con outras linguas con maior número de falantes, e tamén coas outras linguas co-oficiais do estado español (catalán e éuscaro). Esta diferenza é crítica en canto a recursos e servizos relacionados con datos de tipo multimedia ou do ámbito da saúde, pois os existentes son pobres en diversidade e pequenos en tamaño. O maior perigo a curto prazo é que de non reverterse esta situación posiblemente o galego quede fóra da revolución que o Big Data e a Intelixencia Artificial está a provocar en moitos sectores estratéxicos polo simple feito da falta de recursos para aplicar estas tecnoloxías.

Por outra banda, é salientable a existencia dunha experimentada comunidade investigadora en áreas tales como o recoñecemento automático do fala, a síntese de voz ou o procesamento de linguaxe natural e, por suposto, en áreas humanísticas como a filoloxía ou a lingüística. Un feito interesante é a colaboración interdisciplinar entre ambas as áreas, tanto para o desenvolvemento conxunto de ferramentas e recursos lingüísticos, como para investigacións puramente teóricas. Esta comunidade é en gran medida a responsable nos últimos

catro anos dun aumento considerable da produción de recursos e servizos de calidade para a lingua galega.

Outro dato interesante resultado desta investigación é que a pesar de que a industria galega baseada en tecnoloxías da linguaxe é escasa, a existente posúe unha gran compoñente de base tecnolóxica proveniente de spin-offs de universidades públicas e centros de investigación galegos. Este dato fálanos dunha boa comunicación entre as entidades produtoras de coñecemento e o tecido empresarial galego. Con todo, é posible deducir das páxinas webs oficiais destas empresas e de documentos oficiais da Xunta de Galicia que os esforzos están se a centrar máis en desenvolver solucións para a lingua oficial do estado, o español, que para o galego. Doutra banda, observamos unha tendencia nas grandes empresas do sector consistente en reducir os seus esforzos en desenvolver tecnoloxías específicas para linguas minoritarias como o galego.

Polo que respecta ao status legal, o galego como lingua co-oficial da Comunidade Autónoma de Galicia está protexido e lexitimado dentro da estrutura do estado español. Ademais, conta con entidades como a Real Academia Galega, o Consello da Cultura Galega, a Mesa pola Normalización Lingüística ou a Asociación PuntoGal, que velan por aspectos formais e de presenza da lingua galega nos espazos virtuais e físicos. Con todo, notamos unha falta de interese polas linguas co-oficiais en xeral dentro dalgunhas estratexias nacionais relacionadas coas tecnoloxías do fala como na Estratexia Procesamento da Linguaxe Natural 2020 (Gobierno_de_España, 2020b) ou a Estratexia Nacional de Intelixencia Artificial 2020 (Gobierno_de_España, 2020a).

Cremos pertinente, a partir da experiencia e os datos adquiridos durante a elaboración deste informe, deixar algunhas recomendacións. En primeiro lugar, sería conveniente a creación dun ente público que se encargue de custodiar de maneira centralizada e estandarizada todos os recursos desenvolvidos para a lingua galega. Este sería un primeiro paso de vital importancia para dinamizar tanto a produción como a distribución de recursos lingüísticos para o galego, pois actualmente atópanse diseminados en páxinas web ou en servidores internos dos desenvolvedores, facendo complexa a súa procura e seguimento. En segundo lugar, é necesario investir na creación de bases de datos de recursos lingüísticos de calidade e gran tamaño con contido multimedia, dígase gravacións de audio ou vídeo, que cubran as distintas variantes e estilos da lingua falada no territorio galego. Unha terceira recomendación sería apoiar a produción científica e a transferencia tecnolóxica baseada en tecnoloxías da fala e a linguaxe para que o galego gañe presenza en solucións comerciais e sexa considerado como un nicho de mercado de interese. En canto ás comunidades de falantes cremos indispensable que consuman e produzan contidos en galego; pero tamén que esixan soporte para o galego naqueles servizos e produtos que consomen regularmente (plataformas de contidos, medios dixitais, aplicacións móbiles ou de escritorio, sistemas operativos, etc.).

Os datos obtidos neste estudo deixan clara a existencia dunha comunidade galega científica e tecnolóxica capaz e interesada na creación de tecnoloxías lingüísticas para a súa lingua, pero con insuficientes recursos como para levala a niveis de soporte e presenza como a do español ou outras linguas co-oficiais. É por tanto vital facer un esforzo substancial e crear recursos lingüísticos para o galego, especialmente de tipo multimedia, como paso imprescindible para acadar a igualdade multilingüe dixital no espazo europeo.

A estrutura do resto deste documento comeza cunha breve introdución, seguida da sección dúas, onde comentamos algúns datos xerais e formais sobre o galego e a súa comunidade de falantes na era dixital. A sección tres presenta unha breve introdución ao campo das tecnoloxías da linguaxe, as súas principais áreas de aplicación/investigación e metodoloxías. Na sección catro, presentamos unha descrición cualitativa de alto nivel dos recursos, proxectos, iniciativas e axentes involucrados nas tecnoloxías da linguaxe para o galego. A sección cinco ofrece unha comparación entre o galego e outros idiomas europeos utilizando métricas desenvolvidas polo proxecto ELE. O documento termina cunha sección de resumo e conclusións.

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 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, experts in more than 30 European languages have conducted an enormous and exhaustive data collection that provided a detailed, empirical and dynamic map of technology support for our languages.[1]

The report has been developed in the frame of the European Language Equality (ELE) project.[2] 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.

2 The Galician Language in the Digital Age

2.1 General Facts

Galician is part of the Romance family of languages. It is the co-official language in Galicia, an autonomous community located in northwestern Spain. Galicia has over 2,600,000 inhabitants. Approximately 1,926,000 persons are speakers of Galician (I.G.E., 2019a). The Autonomous Community of Galicia, the farthest western area of Asturias, León, and Zamora, and three regions in Extremadura, delimit the Galician-speaking territory. Furthermore, due to the historical circumstances of Galician emigration, there are some other regions in the world with a large concentration of people of Galician origin. There are still large Galician-speaking communities in other regions of Spain (Madrid, Barcelona, the Basque Country, and the Canary Islands), Europe (Portugal, France, Switzerland, Germany, the United Kingdom, and the Netherlands), and America (Argentina, Uruguay, Brazil, Venezuela, Cuba, Mexico, and the United States). The number of speakers outside Spain is unknown due to the variety and complexity of the communities.

The Statute of Autonomy of Galicia – passed in 1981 – recognised Galician as the "own" language of Galicia and the co-official language of the Community. The Linguistic Normalisation Act – passed in 1983 – guarantees and regulates citizens' linguistic rights, especially those related to administration, education, and the media. Under the Linguistic Normalisation Act, the local and autonomic administrations are obliged to write all of their official documents in Galician and to establish the use of Galician in the whole educational system. Table 1 shows in data from the most recent census the frequency of use of the Galician language in Galicia.

Galician is closely related to Portuguese. It is also related to other Romance languages like Spanish or French. Galician uses seven different vowel sounds and nineteen consonant sounds. The Galician alphabet has 23 letters (*a, b, c, d, e, f, g, h, i, l, m, n, ñ, o, p, q, r, s, t, u, v, x,z*) and six digraphs (*ch, gu, ll, nh, qu, rr*). The letters *ç, j, k, w,* and *y* are only used in foreign words. The accent mark (´) is used to mark the accented syllable in polysyllabic words and

[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.

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

Census	A Lot	Fairly	Little
2018	57.59%	30.46%	11.95%

Table 1: How often do Galicians speak Galician? (I.G.E., 2019a)

also as a diacritical mark to distinguish between pairs of words that are differentiated in the spoken language because one is stressed where the other is unstressed, or because one of them has a half-open vowel or an open-mid vowel while the other has the corresponding close vowel. In writing, *é* and *ó* can represent both the open-mid vowels as well as the close vowels. Concerning the word order of the sentences in Galician, the principal pattern used is Subject, Verb, Object. Nevertheless, word order in Galician is almost free, and it is not rare to find clitic elements changing the basic structure. In Galician, the passive voice is not usually used, except for scientific, legal, or literary texts. It is possible to form the passive voice using the auxiliary verb *ser* (to be) and the past participle of the main verb. Galician is a pro-drop language: it is possible to use the conjugation of the verb without the personal pronoun involved that plays the subject role. The orthography in Galician is more transparent than in English but less than in Spanish or Italian.

The three main dialectal areas are: eastern Galician, which includes the dialects spoken outside the Galician administrative area, the most important of which is the Galician spoken in Asturias; central Galician, among which the Mondoñedo and Lugo-Ourense varieties stand out; western Galician, where the dialects of the Fisterra region in the north and of Tui and Baixa Limia in the south stand out. The main dialectal phonetic features are: *gheada* (there exists a fricative phoneme or approximant, either voiceless or voiced, in place of the voiced velar occlusive /g/). The *gheada* is characteristic of western Galician and a large part of central Galician; *seseo* (absence of /θ/ and the presence of /s/ in the positions where /θ/ occurs in common Galician, is characteristic of western Galician. The main morphological features are: in nouns, the ending -án in western dialects, as against the ending -ao and -á in the dialects of the central and eastern areas; the formation of the plural of nouns ending in -n, the ending -óns in the western areas, as against the ending -ós in the central area and -ois in the eastern areas; in verbs, the personal suffix -is for the second person plural (*andais*) in the eastern dialects, as against the suffix -des in common Galician (*andades*). The eastern dialects (especially Galician spoken in Asturias) also have many other peculiarities.

2.2 Galician in the Digital Sphere

The presence of Galician on the Internet is limited to less than 0.1% of websites use it (W3Techs, 2021). Nevertheless, there are some initiatives that try to increase the presence of Galician on the web. PuntoGal (PuntoGal, 2021) is an association that has been in charge of obtaining and managing the .gal Internet domain for the Galician community, currently with 6,179 active domains. Another example is Galipedia (the Galician Wikipedia) which ranks 52nd in the number of articles in Wikipedia (Wikimedia Foundation, 2022).

Table 2 shows that in 2018, 76.95% of Galician homes had an internet connection (817,272 homes) and 77.72% of Galicians claim to have used the internet in the three months prior to the survey (I.G.E., 2019b). However, this percentage increases above 98% between the ages of 15 and 44.

An important group of digital content in the Galician language is generated by public institutions of the Autonomous Community of Galicia. The website of the “Corporación Radio e Televisión de Galicia” is an example of multimedia content production. The web also offers a growing number of digital local newspapers in Galician (or Spanish newspapers with a plug-in tool for translation into Galician). Although some digital platforms (Facebook, Youtube)

Age	Man	Woman	Total
5-14	85.68%	85.26%	85.47%
15-24	99.42%	98.82%	99.13%
25-34	99.58%	98.88%	99.23%
35-44	97.95%	98.67%	98.31%
45-54	93.58%	96.00%	94.81%
55-64	82.36%	85.13%	83.79%
65 and more	37.50%	30.65%	33.61%
Total	79.87%	75.72%	77.72%

Table 2: People who used the Internet in the last 3 months (I.G.E., 2019a)

or large software companies (Microsoft, Apple, Google) offer a version with support for Galician in their visual interfaces, many others do not (TikTok, Twitch, Adobe) or use beta version with lower support (Twitter). An extreme lack of support for Galician occurs in the virtual assistants market where none of the four great solutions Alexa, Siri, Google Assistant, or Cortana allow interaction using this language.

A number of products and services have been developed in the last number of years aimed at incorporating Galician to the ICT society. Interesting examples are the web portal of the "Real Academia Galega" (Royal Galician Academy) and the Gaio translator offered as a free web service by the local government of the Autonomous Community of Galicia.

3 What is Language Technology?

Natural language[3] 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 both, 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 intelligent 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 linguistic resources, such as annotated corpora, thesauri, etc. which were manually labelled for various

[3] 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).

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 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). A very popular application 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.[4]

4 Language Technology for Galician

In 2012 META-NET produced a series of white papers about the state of European languages in the digital age (META-NET). One of these white papers was about Galician (García-Mateo and Rodríguez, 2012), and its results were moderately optimistic about the state of LT support for this language. The study concluded that despite an LT community of researchers and a series of state-of-the-art resources and technologies, the scope of resources and the variety of available technologies was very limited compared to the resources and tools for other languages such as Spanish. It was concluded that the Galician LT industry was very small, and it was proposed as the only possible alternative to reverse this situation to make a significant effort to create more and better LT resources for Galician. Ten years later, the LT status for the Galician has changed a bit. We noticed, in our analysis, an increase in the resources and corpora created between 2018-2021 (67.69% of those indexed). However, tools and services developed in the same period have not increased to the same degree (37.27% of those indexed). There is a significant imbalance in the distribution of resources and corpora by technologies. Table 3 shows that corpora for text resources are the most prevalent, whereas corpora for other technologies are very few.

Only text	Multimodal	Only audio	Only video+audio	Only video
91.93%	5.66%	1.61%	0.8%	0%

Table 3: Distribution of corpus and resources base on the media type

Most of the resources come from three types of sources: non-Galician universities and research centers (42.17%), Galician public institutions (28.92%), and non-Galician private companies or public institutions (28.92%). It is important to note that most of the resources, services, and tools created by non-Galician entities tend to belong to multilingual projects or products that include Galician as one of several languages. However, most of the resources, services, and tools created by Galician entities tend to focus on Galician, offering quality in each.

Regarding the accessibility and use of resources for Galician, since most of them have been developed by open-source projects, study centers and universities, they can be downloaded and used under licenses that are mostly compatible with GNU/GPL. However, around 20%

[4] 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).

Tools & Services		
Without a fee for all uses 71.11%	Without a fee for non commercial uses 26.67%	With a fee 2.22%
Corpus & Resources		
Without a fee for all uses 75.44%	Without a fee for specific uses 10.52%	With a fee 14.04%

Table 4: Distribution of licenses by fee

of the indexed elements are not available for commercial purposes. Table 4 shows in detail the distribution of licenses according to their restrictions.

4.1 Language Data and Tools

The situation of Galician in terms of data and resources is optimistic for most of the technologies that process and use text. However, regarding multimedia data, there is an enormous gap and speech processing technologies seem to be less mature than technologies based on text processing.

For Galician, key results regarding technologies and resources include the following:

- There are large reference text databases in the modern and historical Galician with a balanced mix of various domains (Piñeiro, 2019; García-Mateo et al., 2014). There are also corpora specialised in economics, technology, or the legal field.
- There are some databases annotated with syntactic, semantic, or discursive information. However, the number and size of these resources decrease as more complex linguistic and semantic information is needed. This fact could put a brake on text-based technologies, such as text summarisation or text generation on Galician.
- Parallel databases with millions of tokens exist between Galician and other languages such as Spanish, Portuguese, French, and English (Tiedemann, 2012). These databases have been used to develop machine translation systems quite successfully for nearby languages such as Portuguese or Spanish. However, there is still a lack of data and effective translation systems for other languages.
- A relevant model to highlight is Bertinho (Vilares Calvo et al., 2021). Bertinho is a monolingual BERT model (Devlin et al., 2018) for Galician, with better performance than the well-known official multilingual BERT model (mBERT). Bertinho implements state-of-the-art technology, and it is possible to use it in many NLP tasks as POS-tagging or Punctuation Restauration. However, its developers declare that both with regard to the volume of training data and performance Bertinho still does not reach other monolingual versions, such as BETO for Spanish.
- The multimedia data available is small (the maximum recorded duration is approximately 33 hours), with little domain variability (mainly broadcast), generally consisting of spontaneous voice recordings or phrase readings but with excellent acoustic quality. The amount of multimedia data available makes it a demanding challenge to build state-of-the-art systems based on deep learning for the Galician. Therefore, speech processing technologies such as text-to-speech (TTS) or automatic speech recognition (ASR) in Galician are far from the performance achieved in languages such as English or Spanish.

- Another important gap detected is in the area of human-computer interaction where the necessary tools and resources to build chat-bots, virtual assistants, and similar systems are poor or outdated.

4.2 Projects, Initiatives, Stakeholders

Galician is one of the co-official languages of Spain and the language of the Autonomous Community of Galicia. Spain has national plans for both Artificial Intelligence (Gobierno_de_España, 2020a) and Language Technologies (specifically for NLP) (Gobierno_de_España, 2020b). These national plans focus more on the potential, opportunities, and needs of the Spanish's LT, giving less importance to co-official languages such as Galician. Two national associations bring together the community of researchers on issues related to LT: Sociedad Española de Procesamiento del Lenguaje Natural with focus on NLP, and the Red Temática en Tecnologías del Habla with its focus on speech processing.

The Autonomous Community of Galicia has its own strategy for AI (Xunta_de_Galicia, 2021). This document describes the current environment of AI in Galicia and provides a roadmap for public investments and developments until 2030. According to this report, by 2021, there are about 258 projects related to AI in the Galician ICT environment (13 of them refer to NLP and 9 to cognitive assistants). However, there are many more projects related to LT in the Galician university environment, both from a linguistic and technological point of view. Another interesting fact is that from the number of companies in the Galician ICT industrial environment that use AI, only 21% are focused on cognitive assistants and just 12% on NLP (Xunta_de_Galicia, 2021). The Galician LT industry is very small, but a very active environment of spin-offs and public programs exist dedicated to transferring knowledge from universities to the market.

5 Cross-Language Comparison

The LT field[5] 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:

- The current state of technology support, as indicated by the availability of tools and services[6] 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)

[5] This section has been provided by the editors.

[6] 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.

  - 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[7]

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

[7] 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[8] 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.[9]

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 5 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,[10] 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

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

[9] 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.

[10] In addition to the languages listed in Table 5, 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.

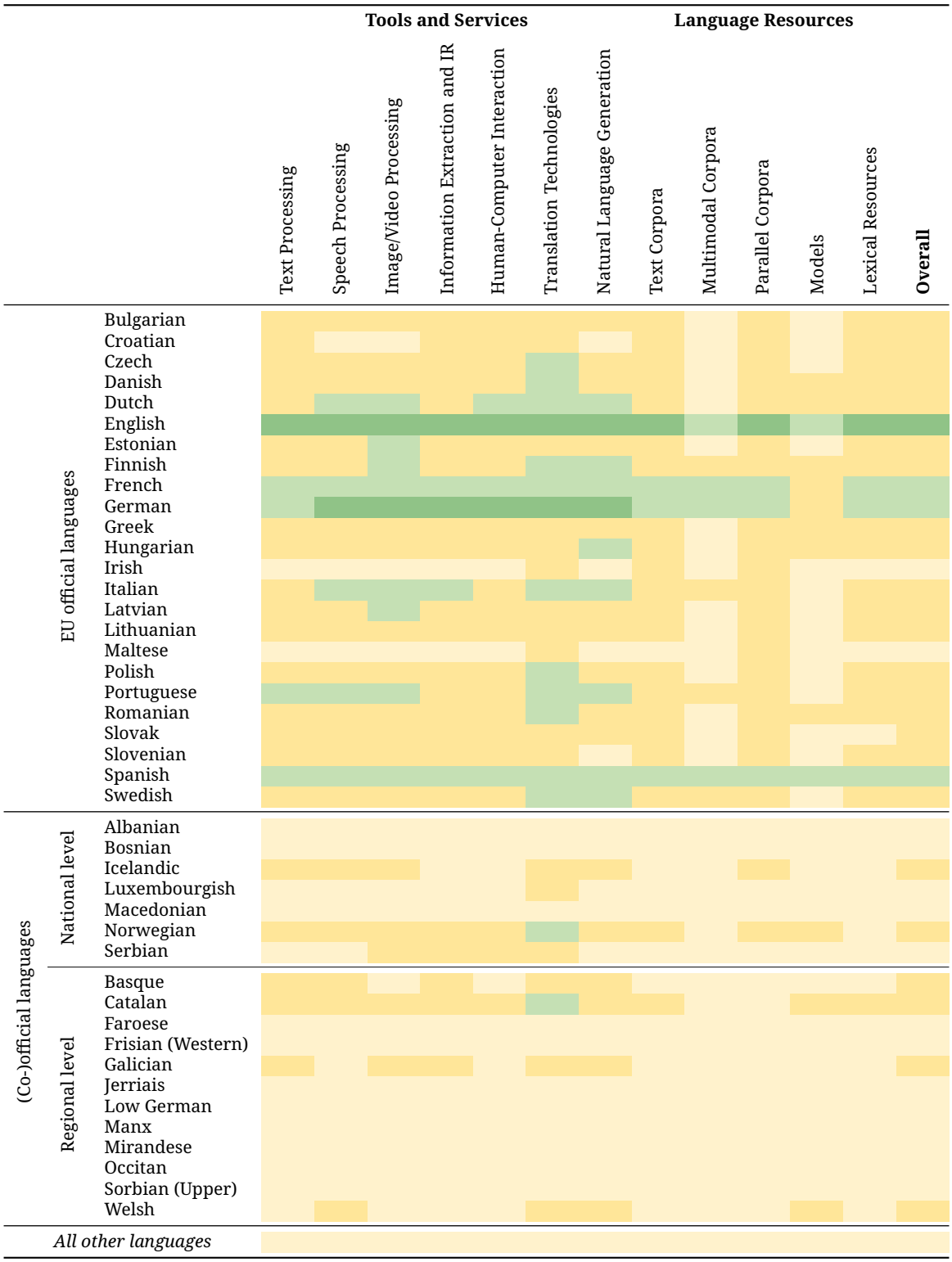

Table 5: State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support)

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

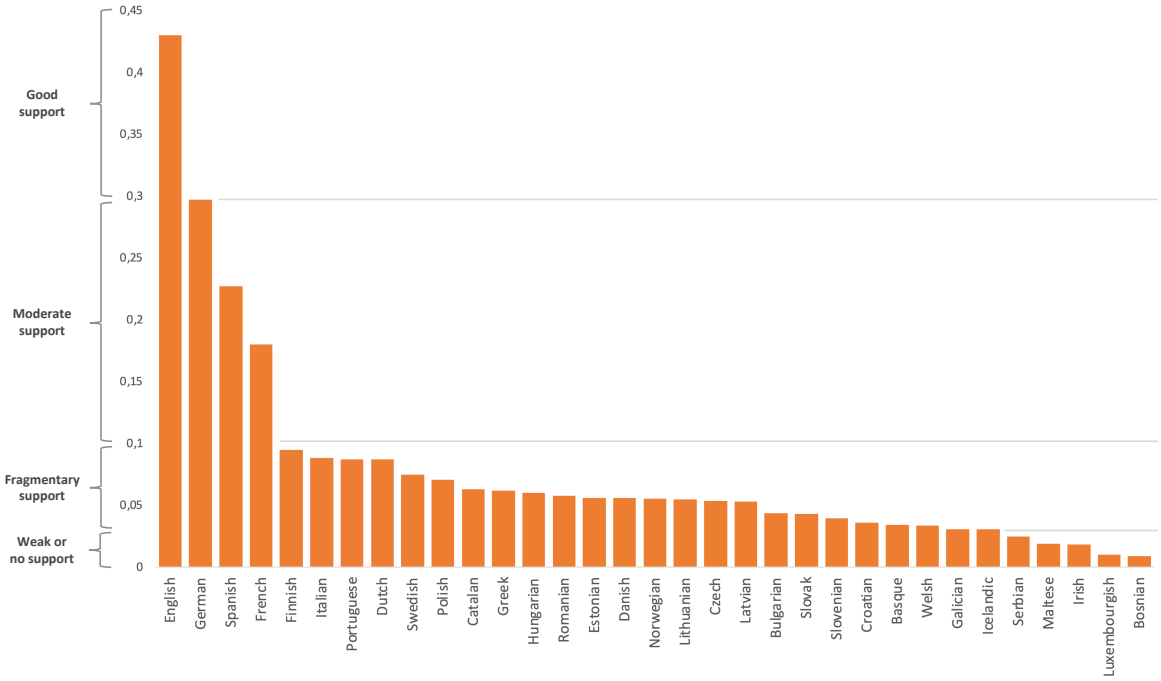

Figure 1: Overall state of technology support for selected European languages (2022)

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

This report shows that there are huge differences between European languages. In the case of Galician, there are substantial gaps in both resources and tools, especially in those based on multimedia data. There are national and regional plans to invest and stimulate the development of LT but they focus on the official language of Spain and have less focus on Galician. We also notice a strong LT research community in Galicia supported by national and local research programs, and impressive growth in the amount of data and resources created in the last four years. However, the scope of the resources and the range of tools are still limited compared with the number of resources and tools available for other languages such as English or Spanish, and they are not sufficient in terms of quality or quantity to develop state-of-the-art technologies based on data greedy paradigms such as deep learning models. There are a few specialised medical datasets but of small size, poor quality, and all of them are text datasets.

The Galician LT industry is currently very small but with a high spin-off component and good knowledge transfer between universities and companies. International companies have either stopped or severely cut their LT efforts for Galician, usually using automatic translation technologies.

Our report shows the necessity to make a substantial effort to create LT resources for Galician, especially multimedia resources, to provide a multilingual digital equality space in Europe. The need for large amounts of data is now more urgent than ever due to the great potential that artificial intelligence and big data can offer. These technologies are already crucial today, and there is a danger for under-resourced languages like Galician to be left behind in the future.

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.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Carmen García-Mateo and Montserrat Arza Rodríguez. Language technology support for galician. In *The Galician Language in the Digital Age*, pages 50–67. Springer, 2012.

Carmen García-Mateo, Antonio Cardenal López, Xosé Luis Regueira, Elisa Fernández Rei, Marta Martinez, Roberto Seara, Rocío Varela, and Noemí Basanta. Corilga: a galician multilevel annotated speech corpus for linguistic analysis. In *LREC*, pages 2653–2657, 2014.

Gobierno_de_España. Estrategia nacional de inteligencia artificial 2020, 2020a. URL https://portal.mineco.gob.es/RecursosNoticia/mineco/prensa/noticias/2020/201202_np_ENIAv.pdf.

Gobierno_de_España. Estrategia procesamiento del lenguaje natural 2020, 2020b. URL https://drive.google.com/file/d/1eXlFdRNTmOx4sm3FQ439Z8zaeNqEFGiK/view.

I.G.E. Enquisa estrutural a fogares. coñecemento e uso do galego. resumo de resultados 27/09/2019, sep 2019a. URL http://www.ige.gal/estatico/estatRM.jsp?c=0206004&ruta=html/gl/OperacionsEstruturais/Resumo_resultados_EEF_Galego.html.

I.G.E. Enquisa estrutural a fogares. novas tecnoloxías. resumo de resultados 31/07/2019, 07 2019b. URL http://www.ige.eu/estatico/estatRM.jsp?c=0205002&ruta=html/gl/OperacionsEstruturais/Resumo_resultados_EEF_NovasTecnoloxias.html.

META-NET. meta-net.

Centro Ramón Piñeiro. Corpus de referencia do galego actual (corga) [3.2], 2019. URL http://corpus.cirp.gal/corga/.

PuntoGal. Puntogal association, 2021. URL https://dominio.gal.

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.

Jörg Tiedemann. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

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

David Vilares Calvo, Marcos García González, and Carlos Gómez Rodríguez. Bertinho: Galician BERT representations. *arXiv preprint arXiv:2103.13799*, 2021.

W3Techs. Usage statistics of galician for websites 07/12/2021, 12 2021. URL https://w3techs.com/technologies/details/cl-gl-.

Inc. Wikimedia Foundation. List of wikipedias, Jan 2022. URL https://meta.wikimedia.org/wiki/List_of_Wikipedias.

Xunta_de_Galicia. Estratexia galega de intelixencia artificial 2030. cara a unha galicia intelixente, 2021. URL https://amtega.xunta.gal/sites/w_amtega/files/20210608_estrategia_ia_gl.pdf.