EUROPEAN LANGUAGE EQUALITY

D1.6

Report on the Catalan Language

Authors	Maite Melero, 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	D1.6
Deliverable title	Report on the Catalan Language
Type	Report
Number of pages	24
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, Blanca C. Figueras, Mar Rodríguez, Marta Villegas
Reviewers	German Rigau, Teresa Lynn
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
CH	Cultural Heritage
CL	Computational Linguistics
CLARIN	Common Language Resources and Technology Infrastructure
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)
ELRC	European Language Resource Coordination
EU	European Union
GPU	Graphics Processing Unit
HCI	Human Computer Interaction (see HMI)
HMI	Human Machine Interaction (see HCI)
HPC	High-Performance Computing
IEC	Institute for Catalan Studies
LR	Language Resource/Resources
LT	Language Technology/Technologies
META	Multilingual Europe Technology Alliance
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
SR	Speaker Recognition
STT	Speech-To-Text
TTS	Text-To-Speech

Abstract

This report has been developed in the framework of the European Language Equality (ELE) project,[1] 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 unforeseeable before. The wide scope of language applications evidences demonstrate 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.

It is unquestionable that the situation of the European languages in the LT field is very unequal, to the point of compromising the future of the lesser-spoken languages. LT in a given language is crucially dependant on the availability of high quality language resources in the form of large amounts of text and speech data, annotated corpora, lexicons, benchmarking tools, etc. It is a fact that the most advanced AI products exist only for a handful of languages, leaving speakers of minority languages out of the technological progress. 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, all the more so for non-global languages such as Catalan, which is the object of this report. In this report, we provide a snapshot of the current situation of LT in Catalan, based on a comprehensive survey of language resources and tools, which have been collected and documented in the European Language Grid (ELG),[2] where further details can be consulted and the resources accessed. While the present situation presents a mixed picture, with many important gaps still to fill, it is true that the recent AI revolution has resulted in an increased awareness of the Catalan society and political bodies, of the importance of LT, as evidenced by the recent launch of the AINA project,[3] aiming at building resources for Catalan. This increased awareness should hopefully lead to a more mature, comprehensive, sustainable LT ecosystem for the Catalan language.

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 procedure that provided a detailed, empirical and dynamic map of technology support for our languages.[4] The reports have been developed in the framework of the European Language Equality (ELE) project.

[1] https://european-language-equality.eu
[2] https://www.european-language-grid.eu
[3] https://www.projecteaina.cat
[4] 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.

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 Catalan language and gives a snapshot of the current situation of LT in Catalan, which is a moving target due to the rapid advances of AI-based technologies. These advances, in the context of unstoppable globalisation, put a lot of pressure on non-global languages, such as Catalan, in serious danger of becoming digitally extinct. Having access to technological resources for Catalan is an important factor for survival in the digital era. In the midst of the current AI revolution, Catalan society and political bodies are gaining awareness of the importance of language technologies and resources as shown by the recent publicly-funded AINA project, aiming at building resources for Catalan, and the success of crowd-sourcing campaigns like the Common Voice initiative.[5] This newly acquired awareness should hopefully lead to a better funded, sustainable language technologies ecosystem for the Catalan language.

2 The Catalan Language in the Digital Age

2.1 General Facts

Catalan is a Romance language of the Occitano-Romance branch spoken in four European states: Andorra, Spain, France and Italy, where it shares space with three big languages (Spanish, French and Italian). Andorra is the only territory where Catalan is the only official language. In Spain, it is mainly spoken in Catalonia, Valencia, and the Balearic Islands, where it is official together with Spanish, and in two small territories in Aragon (Franja de Ponent) and Murcia (the Carxe). In Valencia, the traditional denomination of the language is Valencian. Catalan is also official in Alghero together with Sardinian and Italian. Dialectal variation in Catalan can be grouped in two main blocks, mostly based on phonetic and morphological differences: Eastern Catalan and Western Catalan.

Figure 1: Catalan-speaking area. Source: Institut Ramon Llull

Table 1 shows the percentages of Catalan speakers relative to the population of the different territories it is spoken and the total number of speakers. According to this number, Catalan ranks 13th in the languages spoken in the European Union. Studies show that the use

[5] https://www.projecteaina.cat

of Catalan as an everyday language is receding (Sorolla and Vila i Moreno, 2018). As an example, in Catalonia, the most populated of all the Catalan-speaking territories, this percentage went from 46% in 2003 to 36.1% in 2018, mostly due to inbound migration.

	Habitual speakers (%)	Habitual speakers	Officiality
Catalonia	36.1 (2018)	2,784,092	co-official
Valencia	28.1 (2015)	1,237,659	co-official
Balearic Islands	36.8 (2014)	431,128	co-official
Northern Catalonia	1.3 (2015)	6,198	non-official
Andorra	37.9 (2018)	29,466	official
Franja de Ponent	49.6 (2014)	25,288	non-official
Alghero	9.1 (2015)	3,875	co-official
Total	32.4	4,517,706	

Table 1: Habitual speakers of Catalan in the traditional Catalan-speaking territories. Source: InformeCAT 2021

In Catalonia, Catalan is the main vehicular language for university and non-university teaching as recognised by Article 6 of the Autonomy Statute of Catalonia,[6] although in practice teachers can use Spanish if they deem it fit. In Valencia and in the Balearic Islands, Catalan is a compulsory subject at all levels except university level, although the possibility to do the full schooling in Catalan is also recognised.[7] In Andorra, three different educational systems coexist: Andorran (in Catalan), French and Spanish.[8]

2.2 Catalan in the Digital Sphere

Catalan is sometimes considered a *minoritised language* (Tenedero, 2017) because of its subordination or unequal power relationship with major languages with which it shares territory, namely Spanish, French and Italian. Despite this status of minoritised language and its vulnerable position, in most of the areas where it is spoken, its presence in the digital sphere is relatively strong. A good example of this is the Catalan Wikipedia, which ranks 20th globally in terms of number of articles (Belmar, 2019).

In 2020, the Internet domain *.cat* had 113,391 registered websites, out of which 35.7% had their landing pages in Catalan. According to the November 2021 Catalan Internet Barometer,[9] the overall percentage of use of Catalan in public and commercial websites is high, 66.03%, and has been steadily growing since the Barometer started its observations in 2002, when it was just 38.75%. However, digital presence of Catalan is uneven across sectors: out of the 35 sectors analysed in this study, 10 have a low use of Catalan (below 50%), among them high impact sectors, such as automobile and technology-related multinational companies. In fact, only 30.3% of the 480 most popular brands in Catalonia have their website translated to Catalan. In contrast, universities, NGOs and culture-related organisations have percentages close to 100%.

With regard to public administration websites, the situation varies widely depending on the level of the administration. Virtually all websites of the Catalan administration, including

[6] https://www.parlament.cat/document/cataleg/48089.pdf

[7] 1983's law on the use and teaching of Valencian (https://www.avl.gva.es/documents/31987/97442/Documents_02.pdf), and the article 35 of the Autonomy Statute of the Balearic Islands's (http://web.parlamentib.es/RecursosWeb/DOCS/EstatutAutonomiaIB.pdf)

[8] Govern d'Andorra – A024. Estadística de pre-ensenyament superior, A025. Estadística de formació professional, Curs 2018/2019

[9] https://wiccac.cat/2021/10/barometre-de-lus-del-catala-a-internet-octubre-2021/

municipalities, use the *.cat* domain and have landing pages in Catalan. In contrast, only 33.3% of websites belonging to the Spanish Administration have a Catalan version, most of them linked to the Ministry of Finance. Coverage at the European level is even poorer, with no relevant websites of the European Union offering a Catalan version. As for social media and streaming platforms, popular sites such as Instagram, Netflix, Spotify, HBO, Linkedin or Tik Tok do not offer localised Catalan versions.

Despite the lack of support provided for Catalan by large platforms, Catalan web users are considerably active online. According to statistics gathered by the Global Language Network, Catalan is the 10th EU language (and 19th of the world) in terms of number of tweets, 9th of the EU (and 17th of the world) in terms of number of users who tweet in this language and 5th of the world in number of tweets per user as stated in the InformeCAT 2020 report.[10] In the last ten years, grassroots social-media initiatives have emerged, such as Valençúbers, Canal Malaia or Creators.tv. These efforts have given visibility to more than 500 Catalan content creators on various channels, such as YouTube, Instagram, Twitter, TikTok or Twitch and have generated millions of views. For example, in TikTok, the #EstikTokat hashtag is the most popular hashtag in the Catalan-speaking community: it brings together a large part of the Catalan content and already has over 300 million views.[11]

With respect to large technology companies, they currently do not incorporate Catalan in their most cutting-edge applications, despite the fact that demand exists. None of the current 32 voice assistants in the market (including Apple's Siri, Amazon's Alexa, Microsoft's Cortana, and Google's Assistant) have the capacity to recognise, understand or speak in Catalan. Some, like Microsoft and Google, do offer Catalan in their well-known translators as well as in lower-level services, like for-a-fee services for recognition and synthesis of Catalan text in their cloud platforms.

In sum, the picture of Catalan in the digital domain is bittersweet, we see a marked tendency for big corporations to ignore the Catalan market and subsume it with the general Spanish market and, at the same time, we have a Catalan-speaking community very active on the web, digitally connected and eager to create content in their language.

3 What is Language Technology?

Natural language[12] 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*.

[10] InformeCAT2020 https://www.plataforma-llengua.cat/media/upload/pdf/informecat-2020_267_11_2406.pdf

[11] InformeCAT2021https://www.plataforma-llengua.cat/media/upload/pdf/informecat-2021-web-bo_290_11_2442.pdf

[12] 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).

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 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 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 revolutionizing 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 a 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.[13]

4 Language Technology for Catalan

Language technologies aroused an early interest in Catalonia. From the mid-nineties, machine translation between Catalan and Spanish began to be used intensively for the purposes of producing bilingual press publications and in the Catalan Administration. Among the products developed during those years, FreeLing[14] (a text analysis tool) and the AnCora[15] annotated corpus still stand out. Like most resources for Catalan, they were produced through public funding by research groups in universities and research centers. In addition, Catalan has been traditionally supported by a strong community of free and open source software, best represented by Softcatalà,[16] which has been actively working in favor of the presence of Catalan in the digital world since 1998. Thanks to this community, many applications (such as Firefox, Libre Office or Ubuntu) were quickly adapted to Catalan. The potential for mobilising Catalan speakers in digital initiatives has been repeatedly demonstrated by successful experiences such as Mozilla's international Common Voice initiative, where Catalan is the fourth language in terms of collected speech data.

The META-NET White paper on the Catalan Language in the Digital Age (Moreno et al., 2012), published almost a decade ago, concluded that while the scope of resources and the range of tools available were still very limited at that point, when compared with Spanish or English, the perspective was moderately optimistic based on the existing state of language technology support. While the years that followed the META-NET report were not particularly favourable to the language industry for lesser-spoken languages, recent technological

[13] 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).

[14] Padró i Stanilovsky, 2012 FreeLing 3.0: Towards Wider Multilinguality Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)

[15] http://clic.ub.edu/corpus/es

[16] https://www.softcatala.org

developments such as the transformer-based models, which are driving the expansion of the AI-based language technologies everywhere, are also boosting the Catalan language industry as well.

In the upcoming sections, we will review the resources available for Catalan at the time of writing this report, and the projects and stakeholders that have made them possible, while also highlighting 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)[17], where further details can be consulted and the resources accessed. Note also that at the time of writing, a specific project to generate resources for Catalan has just begun; the AINA project (see 4.3) is expected to have a very favourable impact on the landscape of technological resources for Catalan, and will subsequently make some of the observations in this report outdated. The reader is referred to the aforementioned European Language Grid to find the latest technological developments for the Catalan language.

In this section, we will first review existing language data and resources, and then we'll turn our attention into available tools and applications.

4.1 Language Data

Language data, in the form of text and speech corpora, is the most important resource for building 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 processed by the machine. Current semi-supervised methods require large unannotated corpora to train and smaller, manually annotated corpora for fine-tuning and evaluation.

Monolingual corpora

Current language technologies rely heavily 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. When it comes to large monolingual unannotated textual corpora in Catalan, the very recent CaText (Armengol-Estapé et al., 2021) is the largest one with acceptable quality. It combines several previously existing large datasets (such as DOGC, CaWac, OSCAR, Open Subtitles, Catalan Wikipedia) plus crawlings of the Catalan Government websites, the 500 most popular .cat and .ad domains and a news corpus. CaText has been automatically cleaned and de-duplicated, resulting in a 1760-million-token raw corpus of text.

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. As for corpora with rich linguistic annotation, the aforementioned AnCora is still the largest and more complete. It consists mainly of news texts, and has been annotated at many levels, including lemma and part of speech, syntactic constituents and functions, argument structure and thematic roles, named entities and others. This corpus has 500,000 tokens and has been iteratively re-annotated by different research groups over several years. It is worth mentioning that the Catalan AnCora treebank is part of Universal Dependencies, a framework for consistent annotation of grammar across languages.[18]

Both CaText and AnCora are two major resources for Catalan and are openly available for all uses.

Morphosyntactic tags, like parts of speech and lemmas are the most common annotations, as well as named entities and transcription. Not surprisingly, there are more corpora annotated for traditional NLP tasks, such as Sentiment analysis or Topic detection, than for more

[17] https://www.european-language-grid.eu
[18] https://universaldependencies.org/ca/index.html

recent tasks, such as detection of hate speech or bias. A series of datasets annotated for text classification, question answering, semantic textual similarity, part-of-speech, and named entity recognition have been recently released as part of the Catalan Language Understanding Benchmark (CLUB) (Armengol-Estapé et al., 2021). Among the resources for Catalan documented in ELG, there are a couple of corpora annotated for sentiment, the most relevant being MultiBooked, based on hotel reviews, and 3 more for stance detection. There is also one recently created corpora annotated for hate speech detection (Cyberbullying), albeit not open, and 2 more for summarisation (CaSum and DACSA).

Broad coverage of a language in a corpus requires having data from all kinds of domains and geographical variation. In terms of domain, there are 7 documented corpora in the news and administrative language domain, 7 corpora covering literary texts, 3 corpora from social media, 2 corpora in the education domain, 2 corpora from the health domain, 1 dataset from the legal domain, 1 corpus relating to cultural heritage, and 1 corpus from the technological domain. All of them are freely available for non-commercial purposes. However, with the exception of those from the news and administration domains, the rest are generally quite small.

Regarding geographical variation, there are some relevant corpora corresponding to the Valencian sub-variant, such as the *Corpus Informatitzat del Valencià* and *Corpus Toponímic Valencià*.

All these resources are valuable, but they are not nearly enough to train high-quality models for a diversity of domains, genres, and tasks. Many of these corpora are too small or lack adequate licenses for distribution. Moreover, evaluation and benchmarking of NLP tools requires well-annotated, specialised corpora, which are still lacking, although this is the target of the aforementioned AINA project in the short term.

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. For obvious reasons, most of the largest bilingual corpora are bilingual Catalan-Spanish, but, to our knowledge, some of the most important ones are not publicly available (e. g. bilingual press). Most of the available multilingual corpora involving Catalan have been collected by the OPUS initiative or have been crawled by the Paracrawl project. They involve mostly bilingual or multilingual websites, including those from the Catalan administration. More high-quality parallel corpora are needed to develop machine translation systems between Catalan and other languages, like English, French, many European languages, and other world languages like Chinese, Russian and Arabic, that as of today do not exist. Having access to high-quality machine translation systems between Catalan and many world languages would have a high impact on important aspects, such as business (boosting e-commerce for the Catalan industry), social (better integration of migrants) and cultural (facilitating the diffusion of Catalan audiovisual productions).

More efforts are needed to locate existing silos of parallel data, especially Spanish-Catalan and English- Catalan in the different administrations (European, national, regional and local) and make them available.

Catalan Sign Language Resources

Catalan Sign Language (LSC) is a sign language used by more than 25,000 people in Catalonia, 12,000 of whom are deaf. As is the case with most signed languages, the Catalan sign language is not related with the languages spoken in the same territory. Signed languages can be grouped according to their relationship with other signed languages. The Catalan

Model	Architecture	Training Corpus	Corpus size (tokens)
BERTa	roberta	CaText	1760M
WikiBERT-ca	bert	Wikipedia	236M
JuliBERT	roberta	OSCAR	729M
CalBERT	albert	OSCAR	729M
DeepCatalan	ULMFit	Wikipedia	98M

Table 2: Monolingual Language Models in Catalan

sign language is classified in the family of the French sign language, but the transmission to Catalonia would have happened very early and the current relationship is not very evident. It is estimated an intelligibility of the 70% with the Spanish Sign Language. There is an ongoing project to collect a LSC corpus[19] carried out by the Institute for Catalan Studies (IEC)[20] and the Pompeu Fabra University,[21] and some lexicons and learning resources have also been documented. The current amount of data is still insufficient to develop translators and other technology related with LSC, thus more efforts should be devoted to this sensible area.

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 large multilingual models, such as mBERT and XLM-RoBERTa, have been successfully used for many less-resourced languages, including Catalan, in the past year, up to 5 monolingual language models for Catalan have been published (see Table 2).[22] The largest of these models, BERTa, has been evaluated against the CLUB benchmark, outperforming the multilingual models and proving the need for large monolingual language models in order to reach state-of-the-art performance on several tasks (Armengol-Estapé et al., 2021).

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

Current 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. However, those tasks remain an important process for many applications. There exists a number of toolkits and packages that gather and maintain these tools. Some of the most complete toolkits that include Catalan are Freeling,[23] SpaCy,[24]

[19] https://blogs.iec.cat/lsc/2018/03/06/corpus-de-llsc/
[20] https://www.iec.cat
[21] https://www.upf.edu
[22] All the information related to the models and other resources cited in this report can be consulted in the European Language Grid https://www.european-language-grid.eu.
[23] https://nlp.lsi.upc.edu/freeling/node/1
[24] https://spacy.io

UDPipe,[25] and LIMA.[26] Table 3 shows the processing tasks supported for Catalan by each of these packages. We have grouped all those that only have tools for more specific tasks in the column *Others*.

Tasks	Freeling	SpaCy	UDPipe	LIMA	OTHERS
Tokenisation	✓	✓	✓	✓	✓
Sentence segmentation	✓	✓	✓		
Lemmatisation	✓	✓	✓		✓
Stemming					✓
Morphologic Analysis	✓	✓		✓	✓
NER	✓	✓	✓	✓	✓
PoS-tagging	✓	✓	✓	✓	✓
Word Sense Disambiguation	✓			✓	✓
Semantic Role Labeling	✓			✓	
Dependency Parsing	✓	✓	✓	✓	✓

Table 3: Applications available in language toolkits for Catalan

Aside from these basic language processing tasks, common end-user tasks include spellcheckers, grammar and style-checkers, etc. which can be integrated in most content management systems. There are several available for Catalan, including one for the Valencian variety. Softcatalà,[27] among many other open-source applications, offers several good-quality checking tools.

Translation technologies

As for translation into and from Catalan, there exists several online platforms offering translation services, either free or for a fee. In addition, some open-source initiatives have built downloadable translation models that include Catalan. Apertium, 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 they provide acceptable results. Apertium currently offers translators between Catalan and the following languages: Aragonese, Italian, English, Esperanto, French, Sardinian, Aranese or Occitan, Portuguese and Romanian. As for neural translation models, Softcatalà offers translators between Catalan and a series of European languages: German, English, French, Italian, Dutch, Portuguese and Spanish. In addition, OPUS-MT has models for all the previous language pairs, plus Ukrainian. Most of these models are of modest quality.

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 (UPC) has built over the years a set of noteworthy speech resources,[28] although most of them are not freely available. The most ambitious collective initiative to generate open source speech data resources for all languages is the Mozilla Common Voice

[25] https://cran.r-project.org/web/packages/udpipe/vignettes/udpipe-annotation.html
[26] https://aymara.github.io/lima/
[27] https://www.softcatala.org
[28] https://www.talp.upc.edu/page-resources-lists

project. As previously mentioned, this collection has been very successful for Catalan, currently comprising 6,458 different voices and 1,200 hours of recorded speech. ParlamentParla,[29] also an open source corpus, consists of around 611 hours of parliamentary speeches. Aside from smaller transcribed audio corpora for specific purposes (e. g. prosody, clinical, or non-canonical constructions), there is also the *Corpus de Català Contemporani de la Universitat de Barcelona* (CCCUB) collected with the aim of studying social and geographical variation.

Large international Speech-To-Text (STT) and Text-To-Speech (TTS) providers, such as Nuance,[30] include Catalan in their catalogue, so do the cloud services offered by Google[31] and Microsoft,[32] but not Amazon. Local companies, like Verbio,[33] offer customised solutions involving STT and TTS technologies, such as automatic subtitling. All for a fee. A very recent open-source TTS tool for Catalan is Catotron, developed by CollectivaT[34] using deep learning models.

4.3 Projects, Initiatives, Stakeholders

The relevance of the multilingual industry and machine translation in the Catalan-speaking territories led early on to the emergence of a series of small and medium-sized language companies gathered in the Clusterlingua[35] association, which formed the first nucleus of what is now a rapidly expanding sector. At the same time, a number of research groups dedicated to these technologies soon emerged in different universities and research centers. Many of the open-source resources for Catalan have been produced by publicly funded research groups and public entities, although not all the results of funded projects have ended being open to the general use. More emphasis should be placed on collecting, opening and maintaining resources resulting from public funding.

In Spain, the Secretary of State for Digitalisation, promoter of the National Plan of Language Technologies,[36] underway since 2015, also leads the National Artificial Intelligence Strategy, defined in 2020. The objectives of the Plan include the creation of resources for Spanish and the other languages of Spain, including Catalan.

In Catalonia, the AI strategy (Catalonia.ai)[37] is led by the Department of Digital Policies, which has recently approved the AINA project[38] to promote the development of technological applications in Catalan, in collaboration with the Barcelona Supercomputing Center.[39] AINA aims to create synergies with all the stakeholders, both potential data providers (public administrations, Catalan Audiovisual Media Corporation, Institute for Catalan Studies, etc.), academia, industry and open-source initiatives, for the creation of open-license language resources to facilitate building AI applications in Catalan.

In fact, in order to fully normalise the integration of Catalan into common AI applications, it will not be enough to solely create the necessary resources (data, models, etc.) or to develop the technology. It will also be necessary to find ways to bring these resources and these technological developments to the market.

Among the relevant stakeholders in the language industry for Catalan, there are sizeable number of research groups focused on NLP or speech technologies in universities and re-

[29] https://zenodo.org/record/5541827#.YhYLHejMI-Y
[30] https://www.nuance.com
[31] https://cloud.google.com/speech-to-text
[32] https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/
[33] https://www.verbio.com
[34] https://collectivat.cat
[35] https://clusterlingua.cat
[36] https://plantl.mineco.gob.es/tecnologias-lenguaje/PTL/Paginas/plan-impulso-tecnologias-lenguaje.aspx
[37] https://politiquesdigitals.gencat.cat/en/tic/catalonia-ai/index.html
[38] https://politiquesdigitals.gencat.cat/ca/tic/aina-el-projecte-per-garantir-el-catala-en-lera-digital/
[39] https://www.bsc.es, https://huggingface.co/projecte-aina

search centers across Catalonia and Valencia, although in the last decade they have suffered from a persistent drain of talent due to scarce funding. Among these groups:

- CLiC-Centre de Llenguatge i Computació[40] (Universitat de Barcelona)
- TALN-Natural Language Processing research group[41] and COLT-Computational Linguistics and Linguistic Theory[42] (Universitat Pompeu Fabra)
- TALP – Language and Speech Technologies (Universitat Politècnica de Catalunya)
- Tradumàtica – Tecnologies de la Traducció (Universitat Autònoma de Barcelona)[43]
- Text Mining Unit (Barcelona Supercomputing Center)[44]
- Transducens[45] (Universitat d'Alacant)
- Machine Learning and Language Processing (MLLP) Universitat Politècnica de València[46]

The TERMCAT[47] is a public Catalan entity entrusted with the creation of terminological resources and the standardisation of neologisms. Its online dictionary collection contains more than sixty lexicons from the fields of life sciences, industry, technology, human sciences, and legal and economic sciences. Its API allows for downloading terminological repertoires of general interest in different formats, under Creative Commons licences.

The Institute for Catalan Studies (IEC)[48] is an academic institution which seeks to undertake research and study into "all elements of Catalan culture". Its Philological Section was responsible for establishing the spelling norms that became the foundation of modern written Catalan and are still in use today. Officially the IEC provides standards for the language as a whole: the Philological Section has members from Catalonia, Northern Catalonia (located in France), the Balearic Islands, Valencia, Alghero in Sardinia and the Principality of Andorra. However, Valencia has its own language academy, the Acadèmia Valenciana de la Llengua, which nevertheless formally acknowledges that theirs is one variant of the common language. The IEC website gives access to several online dictionaries and textual corpora.[49]

Regarding open-source initiatives, the already mentioned Softcatalà is one of the most important initiatives in the field. It is a non-profit association whose basic aim is to promote the use of Catalan in computer science, the Internet and new technologies. They were founded in 1998 and work on language tools such as spell-checkers, translation models, synonym dictionaries and multilingual dictionaries. Another relevant contributor to open-source solutions in the field is Col·lectivaT,[50] a non-profit cooperative that provides cultural translation, research and technological services for collaborative and linguistic work, and focuses on speech recognition and TA integration.

[40] http://clic.ub.edu/ca
[41] https://www.talp.upc.edu, https://www.upf.edu/web/taln
[42] https://www.upf.edu/web/colt
[43] https://www.uab.cat/web/estudiar/l-oferta-de-masters-oficials/informacio-general/tradumatica-tecnologies-de-la-traduccio-1096480139517.html?param1=1345695508762
[44] https://temu.bsc.es
[45] https://cvnet.cpd.ua.es/curriculum-breve/grp/es/transducens/428
[46] https://www.mllp.upv.es
[47] https://www.termcat.cat
[48] https://www.iec.cat
[49] https://www.iec.cat/llengua/recursos.asp
[50] https://collectivat.cat

Finally, there are no conferences or journals specifically dedicated to Catalan language technologies but it is worth mentioning the recent creation of the NLP ComuniCat[51] community, which gathers researchers and developers from industry and academia, with a common interest on NLP in Catalan.

5 Cross-Language Comparison

The LT field[52] 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[53] 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.)

[51] https://twitter.com/NLPComuniCat
[52] This section has been provided by the editors.
[53] 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.

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 ≥3% and <10% of the ELG resources of the same type
3. **Moderate support**: the language is present in ≥10% and <30% of the ELG resources of the same type
4. **Good support**: the language is present in ≥30% of the ELG resources of the same type[54]

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

5.3 European Language Grid as Ground Truth

At the time of writing (January 2022), the ELG catalogue comprises of 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[55] 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.[56]

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

[54] 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.

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

[56] 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.

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 4 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[57], 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 other languages are supported by technology either weakly or not at all. Figure 2 visualises our findings.

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 can 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; Piperidis, 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

[57] In addition to the languages listed in Table 4, 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.

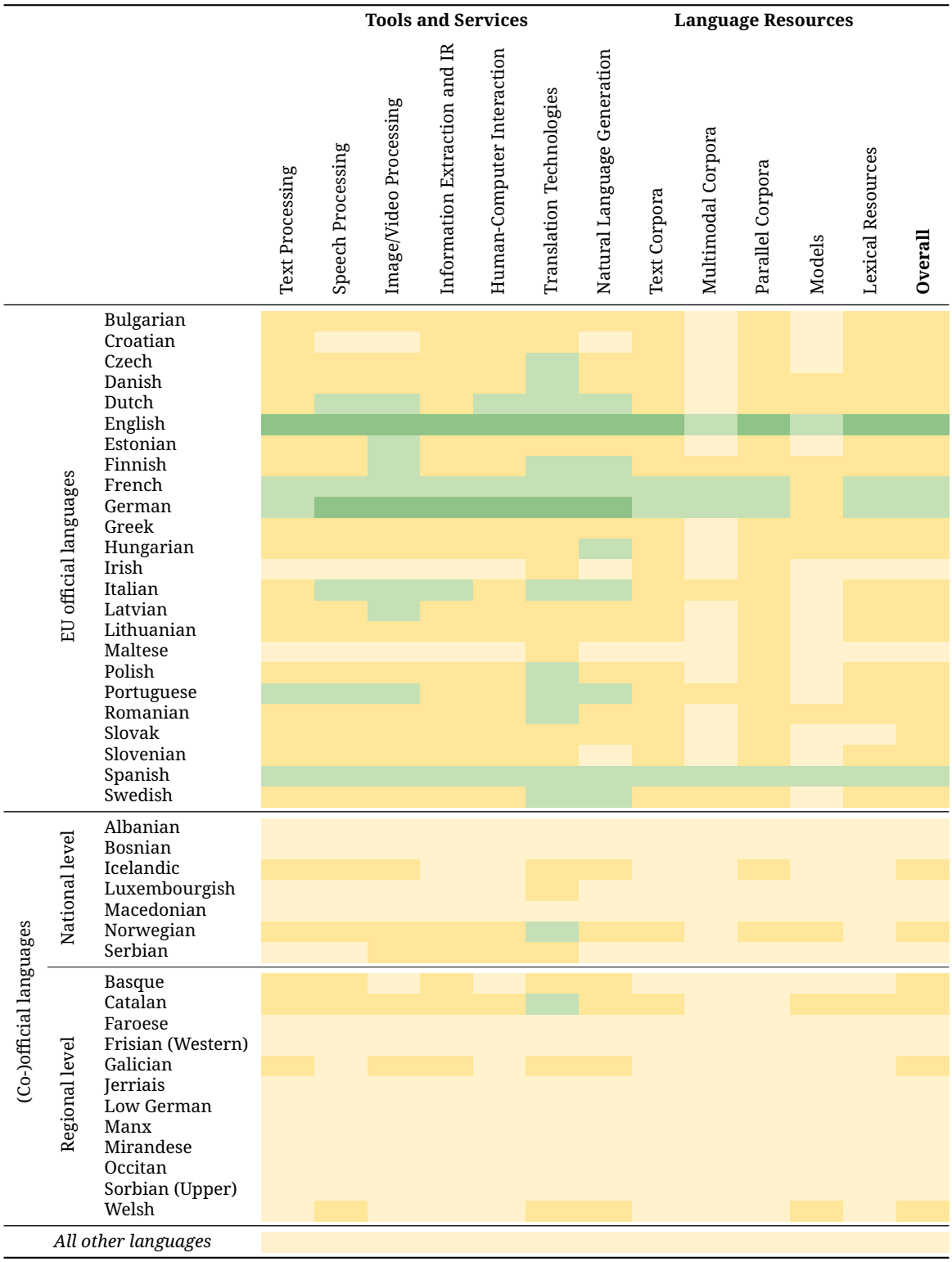

Table 4: State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support)

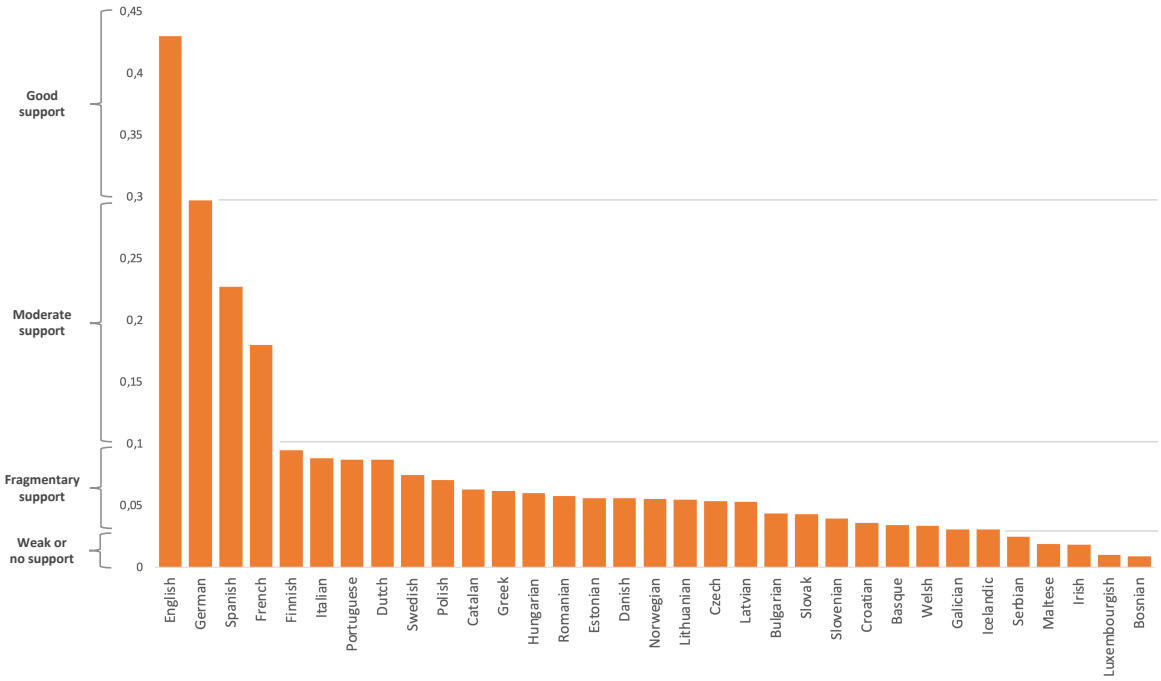

Figure 2: Overall state of technology support for selected European languages (2022)

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 Catalan and provided a snapshot of the landscape at this point in time. The current moment is an evolving situation where AI-based technologies are advancing rapidly, where many languages are trying to catch up with advances for English, in a context of ever-growing globalisation.

The biggest threat that globalisation of Internet content and social networks poses for Catalan, and for thousands of other languages, is digital diglossia. Users in a situation of digital diglossia opt for the language that gives them greater access to content and audience, and allows them to use more advanced technologies. This is true particularly for the younger generations, increasing the generational language gap and bringing the lesser-resourced language to digital extinction. Moreover, in the current technological market dominated by big corporations, another very specific threat to Catalan derives from the fact that the technology giants tend to consider Spain as a single language market, and consequently do not include Catalan in innovative and popular AI applications, such as voice assistants.

On a more positive note, at the level of global opportunities it is worth noting that the latest transformer-based technology, by reducing the need for huge amounts of manually annotated data, thanks to transfer learning and multilingual models, is able to substantially cut down the costs of developing applications, thus making cutting-edge technologies possi-

ble for languages with smaller markets, such as Catalan.

As noted in previous reports, such as the META-NET White Paper on the Catalan Language in the Digital Age (Moreno et al., 2012), the technological profile of the Catalan language community is high and well connected, which accounts for the relatively high presence of the language on the Internet and its dynamism in the creation of new content. In fact, the main strength of Catalan is the mobilisation potential of its community, as shown by the successful collaborative initiatives and the strength of the open source community. Moreover, Catalonia has a strong university and research ecosystem and a public and private sector with long-standing experience in the development and use of language technologies, particularly in machine translation. What has positively changed with respect to the 2012 White Paper is the recent awareness among the public and political bodies of the importance of supporting language resource creation as a way to facilitate inclusion of Catalan in technological products. This is reflected in the recently started AINA project, promoted by the Department of Digital Policies of the Catalan Government.

However, we also find some weaknesses. For a long time there has been a certain reluctance in the Administration to effectively implement the European directives on open data and reuse of public information. For this reason, public administrations and public services still have very large volumes of non-confidential data, suitable for re-use, lying untapped in silos. We expect that with the recent increase in public awareness and projects like AINA, this will begin to change. When it comes to existing resources and applications for Catalan, we have documented a sizeable amount but 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 are lacking the appropriate licenses. Finally, other weaknesses of the sector worth mentioning are the flight of talent abroad (due to lack of opportunity) and the limited academic offer in language technologies. Both circumstances pose a barrier for accessing qualified experts in language technologies, particularly those with an ability to work with Catalan.

Based on this analysis, the following would be appropriate recommendations to ensure the digital future of Catalan:

- An implementation of the European directives on open data to ensure well-regulated open access to language data (text and speech), which is recognised as essential for the development of new products, applications and services in Catalan.
- Support for open source solutions, which will allow small and medium-sized companies (and potentially also large ones) to develop applications in Catalan without having to face the initial investment barrier.
- Increase the innovation capacity of Catalan public services by incorporating cutting-edge technological solutions that include Catalan, thereby acting as a driver of demand in the language technology sector for Catalan.
- Creation of an independent Centre of Excellence dedicated to Catalan Language Technologies, with the aims of (i) increasing visibility and sustainability to the infrastructures and resources, both the existing and the soon-to-be-created by current projects, (ii) offering more educational and training LT programmes in Catalonia to increase the number of trained experts (iii) facilitating technology transfer between academia and industry, (iv) boosting a growing economic sector, while guaranteeing the position of Catalan in the digital challenge.

Acknowledgements

This report has benefited from the insightful comments made by Teresa Lynn 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.

Jordi Armengol-Estapé, Casimiro Pio Carrino, Carlos Rodriguez-Penagos, Ona de Gibert Bonet, Carme Armentano-Oller, Aitor Gonzalez-Agirre, Maite Melero, and Marta Villegas. Are Multilingual Models the Best Choice for Moderately Under-resourced Languages? A Comprehensive Assessment for Catalan. *arXiv:2107.07903 [cs]*, July 2021. URL http://arxiv.org/abs/2107.07903. arXiv: 2107.07903.

Guillem Belmar. LES XARXES VIRTUALS I EL CATALÀ: ACTITUDS, USOS I EL PAPER DE LES COMUNITATS VIRTUALS COM A REFUGIS D'ÚS. *Révue d'Études Catalanes*, (5):26–39, 2019. ISSN 2426-6434. URL https://raco.cat/index.php/REC/article/view/376016. Number: 5.

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

Asunción Moreno, Núria Bel, Eva Revilla, Emília Garcia, and Sisco Vallverdú. *La llengua catalana a l'era digital – The Catalan Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-30677-8. Available online at http://www.meta-net.eu/whitepapers.

Stelios Piperidis. The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 36–42, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/1086_Paper.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. Springer, Heidelberg etc., 2012.

N. Sorolla and F. X. Vila i Moreno. Els grups segons els usos lingüístics i l'evolució de l'ús del català entre 2003 i 2013. *Anàlisi de l'Enquesta d'usos lingüístics de la població a Catalunya 2013*, Vol 1: Coneixements, usos, transmissió i actituds lingüístics:433–460, 2018.

Pia Tenedero. From Minority Languages to Minoritized Languages, November 2017. URL https://www.languageonthemove.com/from-minority-languages-to-minoritized-languages/.

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.