EUROPEAN LANGUAGE EQUALITY

# D1.28

# Report on the Portuguese Language

| | |
|---|---|
| Authors | António Branco, Sara Grilo, João Silva |
| 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 | D1.28 |
| Deliverable title | Report on the Portuguese Language |
| Type | Report |
| Number of pages | 21 |
| 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 | António Branco, Sara Grilo, João Silva |
| Reviewers | Itziar Aldabe, Khalid Choukri |
| Editors | Maria Giagkou, Stelios Piperidis, Georg Rehm, Jane Dunne |
| EC project officers | Susan Fraser, Miklos Druskoczi |
| Contact | European Language Equality (ELE)<br>ADAPT Centre, Dublin City University<br>Glasnevin, Dublin 9, Ireland |
| | Prof. Dr. Andy Way – andy.way@adaptcentre.ie |
| | European Language Equality (ELE)<br>DFKI GmbH<br>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 |
| CAPES | Coordination for Advancement of High Education Personnel |
| CEF | Connecting Europe Facility |
| CL | Computational Linguistics |
| CLARIN | Common Language Resources and Technology Infrastructure |
| CLUL | Center of Linguistics |
| CNPq | National Council for Scientific and Technological Development |
| CRPC | Reference Corpus of Contemporary Portuguese |
| ELE | European Language Equality *(this project)* |
| ELG | European Language Grid (EU project, 2019-2022) |
| ELRC | European Language Resource Coordination |
| EU | European Union |
| FAPESP | Sao Paulo Research Foundation |
| FCT | Foundation for Science and Technology |
| FINEP | Funding Agency for Studies and Projects |
| GPU | Graphics Processing Unit |
| HPC | High-Performance Computing |
| ILTEC | Instituto de Linguística Teórica e Computacional |
| INESC-ID | Instituto de Engenharia de Sistemas e Computadores: Investigação e Desenvolvimento em Lisboa |
| LR | Language Resources/Resources |
| LT | Language Technology/Technologies |
| MCT | Ministry of Science and Technology |
| META | Multilingual Europe Technology Alliance |
| META-NET | EU Network of Excellence to foster META |
| ML | Machine Learning |
| MT | Machine Translation |
| NILC | Núcleo Interinstitucional de Lingüística Computacional |
| NLG | Natural Language Generation |
| NLP | Natural Language Processing |
| NLU | Natural Language Understanding |
| NLX | Natural Language and Speech Group (University of Lisbon) |
| POS | Part-of-Speech |
| SR | Speaker Recognition |
| SVO | Subject Verb Object |

## Abstract

This report provides an analysis of the level of technological preparation of the Portuguese language for the digital age, as well as the actions necessary for the consolidation of Portuguese as a language of international communication with global projection.

The present report on the Portuguese language in the digital age is part of a series of reports that promotes the knowledge about language technology for European languages. It is aimed at the widest possible audience, not specialised in these topics, including language communities, journalists, decision-makers or academics.

This report is part of the European Language Equality (ELE) reports series that provides a detailed empirical survey of the level of technological preparation of European languages. These reports seek to not only to uncover the current state of affairs for each of the European languages covered, but to additionally identify the gaps and factors that hinder further development in Language Technology. Identifying such weaknesses lays the grounds for a comprehensive, evidence-based, proposal of required measures for achieving Digital Language Equality in Europe by 2030.

## Resumo

Este relatório disponibiliza uma análise do nível de preparação tecnológica da língua portuguesa para a era digital assim como das ações necessárias para a consolidação do português como língua de comunicação internacional com projeção global.

O presente relatório sobre a língua portuguesa na era digital faz parte de uma coleção que promove o conhecimento sobre a tecnologia da linguagem nas línguas europeias. É dirigido a um público o mais vasto possível, não especializado nestes temas, incluindo comunidades linguísticas, jornalistas, decisores ou docentes.

Este relatório faz parte da coleção de relatórios da iniciativa European Language Equality (ELE), que fornece um levantamento empírico detalhado do nível de preparação tecnológica das línguas europeias. Esta coleção da ELE procura não só fazer o levantamento do estado atual de cada uma das línguas europeias abrangidas, mas também identificar as lacunas e os fatores que impedem o desenvolvimento da Tecnologia da Linguagem. A identificação de tais pontos fracos permite estabelecer os alicerces para uma proposta abrangente e baseada em dados empíricos das medidas necessárias para alcançar a Igualdade Digital das Línguas na Europa até 2030.


# 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]

[1] The results of this data collection procedure have been integrated into the European Language Grid so that they

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.

Against this background, the present document reports on the technological preparation of the Portuguese language for the digital age.

# 2 The Portuguese Language in the Digital Age

## 2.1 General Facts

Portuguese is the fifth most spoken language in the world, with around 280 million speakers (Instituto Camões in Prague, 2021), of which 250 million are native speakers, spread over four continents: Africa, America, Asia and Europe. It is the official language of Angola, Brazil, Cape Verde, East Timor, Guinea-Bissau, Macau, Mozambique, Portugal, S. Tome and Principe, and Equatorial Guinea.[3]

Due to migratory movements (Pires et al., 2020), Portuguese is also spoken by communities in many countries, occupying an important position in the foreign population in some of them. Camões – Institute for Cooperation and Language, I.P. coordinates Portuguese teaching abroad in several countries, such as Andorra, Australia, Belgium, Canada, China, France, Germany, Luxembourg, Morroco, Namibia, Netherlands, Spain, South Africa, Switzerland, Thailand, United Kingdom, United States of America, Venezuela, and Vietnam (Instituto Camões, 2021).

Portuguese is an official language of the European Union, the Mercosul and the African Union. With the advancement of the alphabetisation in the African countries and in East Timor, Portuguese is confirming its growth potential in terms of the number of speakers.

The expeditions and coastal trade that Portugal maintained across the world for several centuries induced linguistic counterparts that are visible today: Portuguese incorporated words from African, Amerindian and Asian languages, but also gave its lexical contribution to many languages in the world and to several pidgins and creoles of the Atlantic, the Pacific and the Indian Oceans (d'Andrade et al., 1999).

Portuguese is a Romance language (Cardeira, 2006), with most of its lexicon being derived from Latin. At different times in its history, it integrated many words from a variety of languages, which in many cases remain among the most frequent ones. From pre-Latin: *barranco* / ravine, *seara* / cornfield, *bruxa* / witch; Germanic: *luvas* / gloves, *bando* / band, *guerra* / war; Arabic: *aldeia* / village, *açúcar* / sugar, *laranja* / orange; African: *batuque* / drum, *inhame* / yam; Asian: *chá* / tea, *biombo* / partition, *bengala* / walking cane; and Amerindian: *cacau* / cocoa, *tapioca* / tapioca. The languages of the populations that Portuguese contacted during their maritime expansion and coastal trade also integrated Portuguese words. For example, in the case of Japanese, the words *bidoro* (from Portuguese *vidro* / glass) and *pan* (from Portuguese *pão* / bread).

To a speaker not knowing Portuguese, the European variant of this language may often sound like a sequence of consonants. This is due to the fact that, differently from the other Romance languages, the Portuguese unstressed vowels are often weakened or even not pronounced. This vowel weakening is a late change in European Portuguese and it did not affect the variety spoken in Brazil, which in this aspect, is more close to the Portuguese as spoken some centuries ago.

can be discovered, browsed and further investigated by means of comparative visualisations across languages.

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

[3] The content of this chapter is based on a general description of the Portuguese language (Branco et al., 2012)

The basic word order in Portuguese is Subject-Verb-Object (SVO) (*ele leu o livro* / he read the book). In certain pragmatic contexts (e. g. emphatic reading), the VSO order can be used (*lês tu o livro* / read you the book) and the OSV or OVS order are possible in constructions termed as marked by linguists (*o livro, ele não leu* / the book, he not read).

Portuguese is a null subject language, that is the subject of the sentence may not be realised by a phonetically overt expression ( _ *li o livro* / [I] read the book). When the subject is paired with a first person inflection, its non-realisation in phonetic terms is the default option. Additionally, there is no expletive pronoun in impersonal constructions ( _ *há um livro sobre esse tema* / [there] is a book on that subject).

The inflection paradigm in Portuguese is much richer than the one of a language like English, for instance, especially in what concerns verbs: a verb with a regular inflection paradigm will have different markers for aspect, tense, mood, person, number or polarity, reaching more than 160 different inflected verb forms, encompassing both simple and complex ones (Branco et al., 2007).

Also, there are two verb inflectional paradigms which do not exist in the other Romance languages and are very frequent in Portuguese: the inflected infinitive and the future subjunctive. The former shares the theme with the non inflected infinitive (e. g. *cantar* / to sing), to which the aspect, tense, mood, person and number markers are adjoined (*para tu cantares* / for you to sing). The inflected forms of the subjunctive future are homonyms to the ones of the non inflected infinitive, except with irregular verbs, and this increases the number of ambiguous forms in the verbal inflection paradigm.

The geographical division of dialects in Portugal (Cintra, 1999) identifies Southern-Central, Northern and Atlantic island dialects. The Northern dialects can be distinguished by the lack of phonological distinction between /b/ and /v/, with prevalence of /b/, the preservation of ancient diphthongs, and the existence of apico-alveolar fricatives. Differences rely at the phonetic/phonological and lexical levels, being all dialects mutually understandable in an immediate fashion.

All variants of Portuguese across the different continents are mutually understandable. Given their large number, it is not feasible to present here an account of the Portuguese language varieties in other countries and territories, including Brazil.

## 2.2 Portuguese in the Digital Sphere

An overview on statistical data about the Portuguese language reveals that it is one of the most used languages on the internet. According to recent estimates, Portuguese is the fifth most common language on the web, being surpassed only by English, Chinese, Spanish and Arabic (Internet World Stats, 2021). This survey shows that about 172 million users are surfing the web in Portuguese and that in last two decades, from 2000 to 2021, it registered an astonishing expansion of 2 167%.

Portuguese is particularly well positioned when it comes to its presence in social networks. A study of 100 million tweets, reported in (Vicinitas, 2018), reveals that Portuguese is the sixth most spoken on Twitter, after English, Japanese, Spanish, Korean and Arabic.

This is in line with the boom of Internet access, especially among the young people. For instance, Brazil has one of the largest numbers of Internet users worldwide (149 million), of which 139 million are Facebook users (Internet World Stats, 2020a). Portugal in turn has around 8 million Internet users, of which 5.8 million are Facebook users (Internet World Stats, 2020b).

We can also observe the reflex of the presence of Portuguese in the digital age through the growth of Wikipedia. The Portuguese Wikipedia (Portuguese: *Wikipédia em português, Wikipédia em língua portuguesa* or *Wikipédia Lusófona*) is the Portuguese language edition of Wikipedia, the free encyclopedia. *Wikipédia em português* started on 11 May 2001. Being the

sixth most accessed website in the world, *Wikipédia lusófona* is the nineteenth most accessed website in Brazil and the sixth most accessed in Portugal. As of January 2022, it is the 18th largest Wikipedia by article count, containing 1,081,997 articles (Wikipedia, 2022a,b). On October 4, 2020, *Wikipédia em português* approved the restriction on the editing of entries only by registered editors, with the principal objective of preventing vandalism, becoming the first Wikipedia to apply this restriction (Wikipedia, 2022b).

The advent of the digital age and the technological shock it induced for the usage and promotion of natural languages, represents a major challenge for the Portuguese language and for its speakers. The scientific study and the technological development of the Portuguese language, and its preparation for the digital age, is thus a most important endeavour in order to secure the citizenship of its speakers in the information society.

# 3 What is Language Technology?

Natural language[4] 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 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

[4] 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).

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.[5]

[5] In a recent report from 2021, the global LT market was already valued at USD 9.2 billion in 2019 and is

# 4 Language Technology for Portuguese

Although a number of sub-areas in the field have been very active, in terms of language resources and technology, overall Portuguese is a less-resourced language when compared to languages from countries with much larger expenditure in this technology, like English. In this section, the current availability of language data and tools for Portuguese is summarised, taking into account the scientific resources publicly available and above all also those at PORTULAN CLARIN[6] (Branco et al., 2020), with the most comprehensive repository of resources specifically dedicated to the Portuguese language.

## 4.1 Language Data

### Monolingual Corpora

- The number of corpora in Portuguese, both in the European and in the Brazilian variants, has grown in recent years.
- Two large monolingual corpora were compiled for Portuguese, but one lacks representativeness, as it covers only one text type (newspaper), and the other is not fully available due to copyright restrictions.
- A de facto standard 1 million word tagged corpus is available together with the respective POS tagger and other processing tools at the morphological level. For less studied varieties of Portuguese, corpora have been compiled during the last years but they still need to receive more attention, namely by enlarging their size and depth of annotation.
- While some corpora have POS annotation and other types of morphological information, syntactically annotated corpora are smaller and their representativeness in terms of a broader range of domains and genres is in great need to be enlarged.
- There are a number of treebanks of top level quality, together with companion parsers trained on them, and it is necessary to devote more effort on their enlargement.

### Bi- and Multilingual Text Corpora

Quality parallel corpora for machine translation with large volume which include Portuguese are essentially the ones made available by EU initiatives and are consequently very limited in terms of text type (e. g. law texts).

### Lexical/conceptual resources

Much more work needs to be dedicated to lexical resources of all types, including ontologies and the expansion of lexica and wordnets, with still a suboptimal size.

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).

[6] https://portulanclarin.net

#### Models and grammars

Language models to support deep neural processing, including the development of large multimodal language models involving Portuguese, are needed, specially those openly available to be used in research and in innovation

## 4.2 Language Technologies and Tools

- Concerning speech technologies, a number of commercial systems exist for both European and Brazilian varieties (for speech recognition, speech synthesis and dialogue management); although Portuguese and Brazilian teams are active in the field, tools and annotated corpora are usually not available and reserved for internal use in the few laboratories working on speech.
- Tools addressing text and discourse processing are few and partial.
- The same applies to other high level processing tools and applications, like for example, summarisation or conversational systems, among many others.

The above considerations on the availability of data and tools for Portuguese clearly indicate the urgent need to direct substantially more efforts to the preparation of Portuguese to the digital age.

## 4.3 Projects, Initiatives, Stakeholders

The activity in language technology for the Portuguese language can be traced back to projects, programs or initiatives carried out in the last decades.

One of the first important programs in this area was EUROTRA, an ambitious Machine Translation project established and funded by the European Commission from the late 1970's until 1994. The participation of Portugal in this project since 1986 was undertaken by ILTEC, specifically created for this purpose and involving mostly researchers from the Universities of Lisbon and Porto. This project had a long lasting impact on the language industries in Europe with Portugal being no exception. EUROTRA promoted a significant starting step for consistently pursued language technology activities in Portugal and for setting up and fostering a Portuguese community of researchers in this area.

Another example of a European key project in language technology was LE-PAROLE, developed in the late 1990's, with the participation of CLUL and INESC-ID. Its main achievement was the building of corpora and lexicons according to integrated models of composition and materials description. For each language, a 20 million word corpus was built with harmonised design, composition and codification, including a 250,000 word tagged subcorpus. Each language lexicon comprised 20 thousand entries with syntactic and morphologic information.

Part of this corpus was enriched and enlarged under the national project TagShare, conducted at the University of Lisbon, in the Department of Informatics (NLX) and in the Center of Linguistics (CLUL), in 2005. This project enabled the development of a set of linguistic resources and software component tools to support the computational processing of Portuguese. The outcome was a 1 million word corpus linguistically annotated and fully verified by experts, the CINTIL corpus (University of Lisbon, 2006), and a whole range of processing tools for tokenisation, morphosyntactic category (POS) tagging, inflection analysis, lemmatisation, multiword lexeme recognition, named entity recognition, etc., in the LX-* collection, which like many other language data and processing tools for Portuguese, are distributed through the research infrastructure PORTULAN CLARIN.[7]

[7] https://portulanclarin.net

On the basis of these tools and resources, which were extended, top-quality, manually verified treebanks, with syntactic and semantic grammatical analysis, and the companion computational grammar and parsers, have been also developed for the CINTIL-* and LX-* collections, in the national project SemanticShare at the Department of Informatics (NLX Group) of the University of Lisbon, which are also available from PORTULAN CLARIN.

The annotation schemes developed in this project became de facto standards for Portuguese in the field of language technology and have been further used, for instance, in the Reference Corpus of Contemporary Portuguese (CRPC). These results were subsequently expanded in another project, the SemanticShare project, where the construction of a treebank, i. e. the annotation of sentences with their syntactic structure, was undertaken.

The Corpus de Extractos de Textos Electrónicos MCT/Público (CETEMPúblico), released in 2000, in turn, is a corpus of about 180 million words from texts of a Portuguese daily newspaper. It is intended primarily to support the development of processing tools for the Portuguese language which need raw texts for their construction and testing. This corpus was created by the project Computational Processing of Portuguese, under a protocol between the Ministry of Science and Technology (MCT) and that newspaper. This project subsequently evolved into Linguateca, a long term project for Portuguese language technology (Fundação para a Computação Científica Nacional, 2011).

In the field of speech processing, it is worth noting the TECNOVOZ project, which started in 2006. This project was directed by INESC-ID and one of its major goals was to foster technology transfer to the business sector, having as partners companies like the public television RTP.

On the industry side, an important contribution to foster a language technology industry in Portugal was the establishment of the international Microsoft Language Development Center, near Lisbon, which lasted from 2005 to 2015.

In Brazil, relevant efforts in language technology support to Portuguese have been also undertaken.

To mention just a few illustrative examples, in the early 1990's, under the DIRECT project, the Bank of Portuguese was created at the Pontifical Catholic University of São Paulo. Since its inception, the Bank of Portuguese has been a source of data for corpus based studies for several projects.

Also worth mentioning is the Summ-it corpus, a corpus built to support the study of summarisation along with the phenomena of anaphoric and rhetorical relations in Portuguese. This resource was developed under the PLN-BR project, by the Núcleo Interinstitucional de Lingüística Computacional (NILC), driven by the University of São Paulo and gathering researchers from seven other Brazilian institutions.

In 2006 – 2010, the FAROL project was developed, with four participating groups and conducted by the Pontifical Catholic University of Rio Grande do Sul, aimed at reinforcing the cooperation links among teams in Brazil, promoting students and researchers interchange and better research quality in natural language processing.

On a par with these programs and projects both in Brazil and in Portugal, it is worth underlining PROPOR as the key focal initiative of a growing international research community working on Portuguese. PROPOR is the major international scientific conference devoted to the computational processing of Portuguese. This is a biennial conference whose location, since 1993, alternates between the two countries.

The above notes cover only a few illustrative examples of projects, programmes and initiatives in language technology addressing the Portuguese language. Although these are part of positive developments for the Portuguese language in recent years, the fact is that there is a large gap with respect to the language technology activity on other more researched languages, for which the development of language resources and technology is far more advanced.

Compared to the level of funding for language technology not only for English but also for

other languages with far less global projection than the Portuguese language, the support for language technology for Portuguese is still very low.

In Portugal, funding for this area comes mainly from the Ministry of Science, Technology and Higher Education, through the Foundation for Science and Technology (FCT). On a par with FCT, the Fundação Calouste Gulbenkian occasionally funds some language technology projects.

In Brazil, funding for research, in general, and for language technology activities, in particular, is still limited and comes mainly from government agencies. The National Council for Scientific and Technological Development (CNPq), the Sao Paulo Research Foundation (FAPESP), the Coordination for Advancement of High Education Personnel (CAPES), and the Funding Agency for Studies and Projects (FINEP) are the four institutions that significantly support research in this country. Some of these agencies have provided also special joint university-industry funding programs.

A landmark for the language technology for Portuguese landscape is the white paper *The Portuguese Language in the Digital Age* (Branco et al., 2012), produced in the scope of the European META-NET initiative.

Among the private sector initiatives in Portugal in the past few years, it is worth mentioning, DefinedCrowd and Unbabel, two US-based startups with a significant presence in Portugal. Though their business delve around language technology activities, namely by the crowd sourcing, respectively, of the annotation of data sets and the edition of machine translated texts, they are not concentrated on the Portuguese language. The Microsoft Language Development Center, in turn, located in Lisbon and dedicated to speech processing and natural language processing in general, was shut down around 2015, after a decade of activity.

As an outcome of the European project ELRI, funded under the CEF umbrella, the Repository for Translation Resources is available, also known as eTradução.[8] which is maintained since 2019 by AMA, the government agency for the digital transformation of the Portuguese public administration. Several of the data sets therein have been fully validated and uploaded to the ELRC-SHARE repository and are distributed also through this platform.

In June 2019, the Portuguese Government presented the national strategy for Artificial Intelligtence, AI Portugal 2030 (Portuguese Government, 2019), to set out challenges and opportunities of the growing AI ecosystem in Portugal.

The major AI initiative specifically addressing the field of Language Technology is the implementation (2017-2021) and operation (2021-) of the PORTULAN CLARIN Research Infrastructure for the Science and Technology of Language.[9] No other national initiatives, including specific calls for project proposals focusing specifically on Language Technology, have taken place in the past few years, and thus no national funding is available specifically for Language Technology other than the funding of PORTULAN CLARIN (Branco et al., 2020).

Against this background, the key initiative to collect data and processing tools for Portuguese Language Technology and to make them available is the PORTULAN CLARIN national infrastructure, which includes also the Portuguese data sets and language processing tools that have been openly distributed elsewhere.

# 5 Cross-Language Comparison

The LT field[10] 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 re-

[8] https://etraducao.gov.pt/pt-pt/
[9] https://portulanclarin.net
[10] This section has been provided by the editors.

sults 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[11] broadly categorised into a number of core LT application areas:
  - Text processing (e. g. part-of-speech tagging, syntactic parsing)
  - Information extraction and retrieval (e. g. search and information mining)
  - Translation technologies (e. g. machine translation, computer-aided translation)
  - Natural language generation (e. g. text summarisation, simplification)
  - Speech processing (e. g. speech synthesis, speech recognition)
  - Image/video processing (e. g. facial expression recognition)
  - Human-computer interaction (e. g. tools for conversational systems)
- The potential for short- and mid-term development of LT, insofar as this potential can be approximated by the current availability of resources that can be used as training or evaluation data. The availability of data was investigated with regard to a small number of basic types of resources:
  - Text corpora
  - Parallel corpora
  - Multimodal corpora (incl. speech, image, video)
  - Models
  - Lexical resources (incl. dictionaries, wordnets, ontologies etc.)

## 5.2 Levels of Technology Support

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

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

[11] 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.

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[12]

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 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 many major LR/LT repositories[13] and, on top of that, more than 6,000 additional language resources 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.[14]

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 1 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

[12] 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.

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

[14] 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.

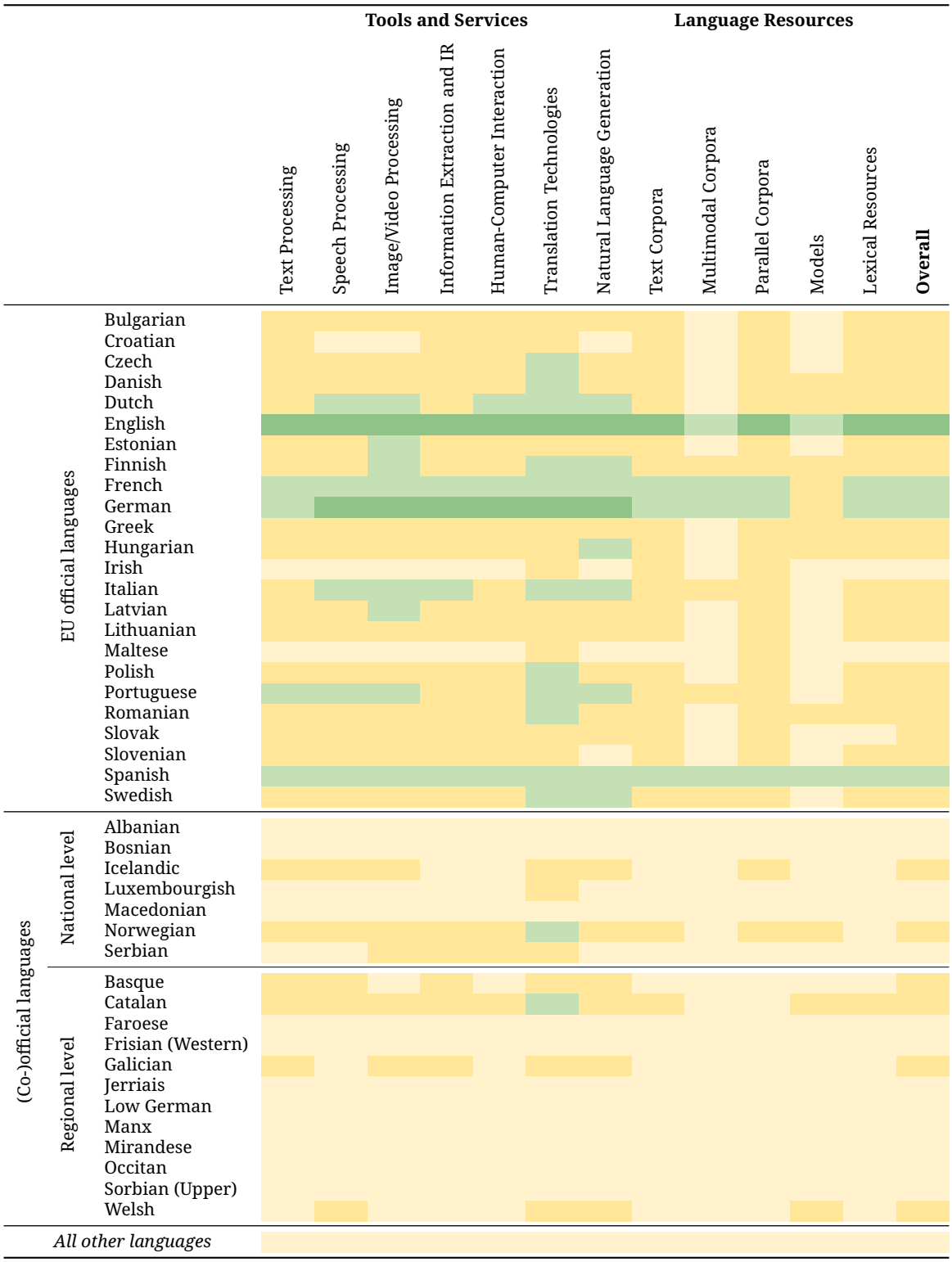


Table 1: State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support)

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[15], 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 1 visualises our findings.

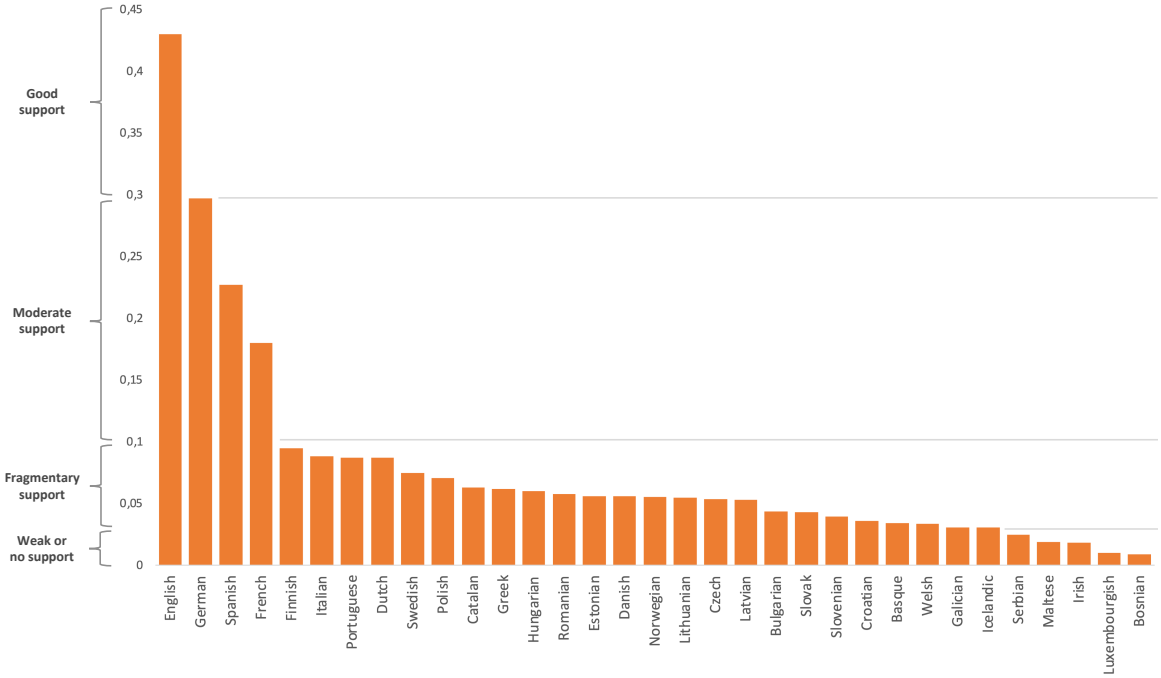


Figure 1: Overall state of technology support for selected European languages (2022)

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 im-

[15] In addition to the languages listed in Table 1, 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.

balance, 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

Overall, the development of Portuguese language data and tools has progressed over the past decade. Although the progress in language technology has also continued at a steady pace, the level of competitive technological preparation of Portuguese for the digital age has not changed significantly over this period when taking the best prepared language, English, as a reference.

Some progress has been made in the area of text analytics and machine translation, thanks to further data collection and corpus creation through a number of initiatives funded by EU-projects and national funds. Fundamental building blocks such as syntactic analysis tools have progressed very significantly, but the underlying datasets still need to be enlarged to build more robust, reliable and application-ready systems.

There is still a large number of fundamental tools and datasets not yet available for Portuguese. While steps have been made towards speech corpus development, there is still no state of the art automatic speech recognition system available for Portuguese as open-source software.

From a natural language understanding perspective, there is a severe lack of semantic-based datasets and tools. Critically, there is a severe lack of freely available, last generation large language models, also known as foundation models, based on deep language learning with artificial neural networks technology.

# 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.

António Branco, Francisco Costa, and Filipe Nunes. The processing of verbal inflection ambiguity: Characterization of the problem space. In *Actas do XXI Encontro Anual da Associação Portuguesa de Linguística*, pages 157–168. Associação Portuguesa de Linguística, 2007.

António Branco, Amália Mendes, Sílvia Pereira, Paulo Henriques, Thomas Pellegrini, Hugo Meinedo, Isabel Trancoso, Paulo Quaresma, Vera Lúcia Strube de Lima, and Fernanda Bacelar. *A língua portuguesa na era digital – The Portuguese Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012. ISBN 978-3-642-29592-8. URL http://www.meta-net.eu/whitepapers/volumes/portuguese.

António Branco, Amália Mendes, Paulo Quaresma, Luís Gomes, João Silva, and Andrea Teixeira. Infrastructure for the science and technology of language portulan clarin. In *Proceedings, 1st International Workshop on Language Technology Platforms (IWLTP 2020)*, pages 1–7. European Language Resources Association (ELRA), 2020.

Esperança Cardeira. *O Essencial sobre a História do Português*. Editorial Caminho, 2006.

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

Luís Lindley Cintra. Nova proposta de classificação dos dialectos galego-portugueses. In *Boletim de Filologia, 22*, pages 81–116. Centro de Estudos Filológicos, 1999.

Ernesto d'Andrade, Dulce Pereira, and Maria Antónia Mota. *Crioulos de Base Portuguesa*. Associação Portuguesa de Linguística, 1999.

Fundação para a Computação Científica Nacional. Linguateca, 2011. URL https://www.linguateca.pt.

Instituto Camões. Co-ordination units for teaching Portuguese abroad, December 2021. URL https://www.instituto-camoes.pt/en/activity-camoes/what-we-do/research/teaching-portuguese-abroad-co-ordination-units.

Instituto Camões in Prague. Português no mundo, 2021. URL https://pt.institutocamoes-praga.cz/centro-de-lingua-portuguesa-instituto-camoes/portugues-no-mundo/.

Internet World Stats. Internet usage, Facebook subscribers and population statistics for all the Americas world region countries, July 2020a. URL https://www.internetworldstats.com/stats2.htm.

Internet World Stats. Internet stats and Facebook usage in Europe 2021 mid-year statistics, December 2020b. URL https://www.internetworldstats.com/stats4.htm.

Internet World Stats. Internet world users by language - top 10 languages, March 2021. URL https://www.internetworldstats.com/stats7.htm.

Rui Pena Pires, Joana Azevedo, Inês Vidigal, and Carlota Moura Veiga. *Emigração Portuguesa 2020*. Observatório da Emigração, December 2020. URL http://hdl.handle.net/10071/21972.

Portuguese Government, editor. *AI Portugal 2030 - Portuguese National Initiative on Digital Skills*. Portuguese Government, 2019. URL https://www.incode2030.gov.pt/en/ai-portugal-2030.

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.

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.

University of Lisbon. Concordanciador CINTIL online (CINTIL online concordancer), 2006. URL https://portulanclarin.net/workbench/cintil-concordancer/.

Vicinitas. 2018 research on 100 million tweets: What it means for your social media strategy for Twitter, March 2018. URL https://www.vicinitas.io/blog/twitter-social-media-strategy-2018-research-100-million-tweets.

Wikipedia. Wikipédia, 2022a. URL https://pt.wikipedia.org/wiki/Wikip%C3%A9dia:P%C3%A1gina_principal.

Wikipedia. Wikipédia em português, 2022b. URL https://pt.wikipedia.org/wiki/Wikip%C3%A9dia_em_portugu%C3%AAs.