Chapter 30
Language Report Portuguese

António Branco, Sara Grilo, and João Silva

Abstract This chapter 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.

1 The Portuguese Language

Portuguese is the fifth most spoken language in the world, with around 280 million speakers (Instituto Camões 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. All variants of Portuguese across the different continents are mutually understandable. 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. This chapter is partly based on Branco et al. (2022) and Branco et al. (2012).

Portuguese has a strong presence in social networks. For instance, a study of 100 million tweets reveals that Portuguese is the sixth most spoken language on Twitter, after English, Japanese, Spanish, Korean and Arabic.[1]

Portuguese is a Romance language, with most of its lexicon being derived from Latin. 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

António Branco · Sara Grilo · João Silva
University of Lisbon, Portugal, antonio.branco@di.fc.ul.pt, sara.grilo@di.fc.ul.pt, joao.silva@di.fc.ul.pt

[1] https://www.vicinitas.io/blog/twitter-social-media-strategy-2018-research-100-million-tweets

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_30

European Portuguese and it did not affect the variety spoken in Brazil, which in this respect is closer to the Portuguese which was spoken some centuries ago.

The basic word order in Portuguese is subject-verb-object (SVO) (*ele leu o livro* / he read the book). Portuguese is a null subject language, where the subject of the sentence may not be realised by a phonetically overt expression (_ *li o livro* / [I] read the book). The inflection paradigm in Portuguese is very rich, especially in verbs. A verb with a regular inflection paradigm will have different markers for aspect, tense, mood, person, number or polarity, culminating in more than 160 different inflected verb forms, encompassing both simple and complex ones.

The advent of the digital age is a major challenge for the Portuguese language and its speakers. The scientific study and technological development of the Portuguese language, making it fit for the digital age, is thus an endeavour of utmost importance in order to ensure that its speakers can participate in the information society.

2 Technologies and Resources for Portuguese

The activity in Language Technology (LT) for the Portuguese language can be traced back to projects, programmes and 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 1970s 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.

Another key European project was LE-PAROLE, developed in the late 1990s, 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. Part of this corpus was enriched and enlarged in the national project TagShare, conducted at the University of Lisbon, in the Department of Informatics (NLX Group) and in the Center of Linguistics (CLUL), in 2005. This project enabled the development of a set of language resources and software tools to support the computational processing of Portuguese. The result was a 1 million word corpus linguistically annotated and fully verified by experts, the CINTIL corpus, 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.

On the basis of these tools and resources, 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. The Corpus de Extractos de Textos Electrónicos MCT/Público (CETEMPúblico), released in 2000, in turn, is a corpus of about 180 million words from excerpts of a Portuguese daily newspaper.

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 fostering an LT industry in Portugal was the establishment of the international Microsoft Language Development Center, near Lisbon, which lasted from 2005 to 2015. More recently, the two US-based startups DefinedCrowd and Unbabel have a significant presence in Portugal.

In Brazil, relevant efforts in LT for Portuguese have also been undertaken. To mention just a few illustrative examples, in the early 1990s, 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, 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.

On par with these programmes and projects both in Brazil and in Portugal, it is worth underlining PROPOR as the key focal initiative of the research community working on Portuguese. PROPOR is the major international scientific conference devoted to the computational processing of Portuguese. The location of this biennial conference has been alternating between the two countries since 1993.

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.

As an outcome of the European CEF project ELRI, the Repository for Translation Resources (eTradução)[2] is available which has been maintained since 2019 by AMA, the government agency for the digital transformation of the Portuguese public administration. Several of its data sets are also distributed through ELRC-SHARE.

The major AI initiative specifically addressing the field of LT is the implementation (2017-2021) and operation (from 2021 onwards) of the PORTULAN CLARIN Research Infrastructure for the Science and Technology of Language.[3]

3 Recommendations and Next Steps

The development of technologies for Portuguese has progressed over the past decade. However, given that progress in LT has accelerated, 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.

[2] https://etraducao.gov.pt/pt-pt/

[3] https://portulanclarin.net

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 entities. Fundamental building blocks such as syntactic analysis tools have progressed significantly, but the underlying datasets still need to be enlarged to build more robust, reliable and application-ready systems.

There are 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 lack of semantic-based datasets and tools. Critically, there is a severe lack of freely available large language models, also known as foundation models, based on deep language learning with artificial neural network technology. Such language models to support deep neural processing, including the development of large multimodal language models involving Portuguese, are thus very much needed, especially those openly available to be used in research and in innovation

The above considerations on the availability of data and tools for Portuguese clearly indicate the urgent need to direct substantially more funding and efforts to the preparation of Portuguese for the digital age. The scientific study and technological development of the Portuguese language is a crucial endeavour for its promotion, in order to ensure that its speakers can participate in the information society.

References

Branco, António, Sara Grilo, and João Silva (2022). *Deliverable D1.28 Report on the Portuguese Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-portuguese.pdf.

Branco, António, Amália Mendes, Sílvia Pereira, Paulo Henriques, Thomas Pellegrini, Hugo Meinedo, Isabel Trancoso, Paulo Quaresma, Vera Lúcia Strube de Lima, and Fernanda Bacelar (2012). *A língua portuguesa na era digital – The Portuguese Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/portuguese.

Instituto Camões (2021). *Português no Mundo*. https://pt.institutocamoes-praga.cz/centro-de-lingua-portuguesa-instituto-camoes/portugues-no-mundo/.