DOI: 10.17469/O2106SLI000011

HELIANA MELLO, TOMMASO RASO, MIGUEL OLIVEIRA, TONY BERBER SARDINHA, CLÁUDIA FREITAS, SANDRA MARIA ALUÍSIO, THIAGO PARDO, MAGALI DURAN, SIDNEY LEAL, MARK DAVIES, CHARLOTTE GALVES-CHAMBELLAND

Brazilian Portuguese: Spoken, Written and Diachronic Corpora

In this paper, a team of corpus linguists who work with Portuguese corpora have joined to report on the major corpora projects dealing with Brazilian Portuguese. All of the corpora are accessible digitally. The projects cover spoken (C-ORAL-BRASIL, Nurc Digital), written (Corpus Brasileiro, Linguateca corpora, NILC corpora) and diachronic corpora (Tycho Brahe). The corpora major characteristics, the projects sites as well as a list of publications pertaining to them are provided.

Keywords: Brazilian Portuguese, spoken corpora, written corpora, diachronic corpora.

1. *Introduction*

In this paper we present sequentially, in the same order that can be seen in the SLI Demo video (https://underline.io/lecture/33029-brazilian-portuguese-spoken,-written-and-diachronic-corpora), the Brazilian Portuguese corpora discussed by presenters during the 2021 SLI conference. The sequence will be as follows: C-ORAL-BRASIL (Mello & Raso), Nurc Digital (Oliveira Jr.), Corpus Brasileiro (Berber Sardinha), Linguateca Corpora (Freitas), NILC Corpora (Aluísio *et al.*), Corpus do Português (Davies) and Tycho Brahe (Galves-Chamberlland).

2. *C-ORAL-BRASIL*

The C-ORAL-BRASIL corpora have been compiled at the LEEL Lab, at the Federal University of Minas Gerais (UFMG), Brazil, by Tommaso Raso and Heliana Mello and their research team.

The C-ORAL-BRASIL corpora portray spontaneous Brazilian Portuguese speech in natural context, media and telephonic interactions. The corpora comprise sound, sound to text alignment, transcription, PoS and parsing files.

The corpora are useful for several types of studies focusing on spontaneous speech, however they were designed so as to be especially adaptable for informational structuring, prosodically based pragmatic studies.

For each corpus, a sample is manually informationally tagged, following the Language into Act Theory (Cresti 2000). The complete collection of corpora encompasses the following:

- C-ORAL-BRASIL I: informal spontaneous speech (Raso & Mello 2012);
- C-ORAL-BRASIL II: formal in natural context speech, media and telephone (forthcoming);
- C-ORAL-ESQ: schizophrenic patients and doctor interactions (forthcoming);
- C-ORAL-ANGOLA: Angolan Portuguese (forthcoming);
- COBAI: Brazilian learners of English speech (2012);
- COLPI: Indigenous Brazilian Portuguese (2016);
- Informationally tagged minicorpora of Brazilian Portuguese, Italian and American English.

The C-ORAL-BRASIL I and II portray an ample range of diaphasic variation, covering private, public, informal and formal monologues, dialogues and conversations, in addition to an assortment of radio and tv programs, besides telephonic conversations. Diastratic variation is well represented, while diatopic variation for the C-ORAL-BRASIL I and the formal and telephonic parts of C-ORAL-BRASIL II are representative of the Belo Horizonte metropolitan area, while the C-ORAL-BRASIL II media part represents the variation that is carried through Brazil's major TV and radio networks.

The C-ORAL-BRASIL materials and products are accessible through the following links:

- Download area and publications: www.c-oral-brasil.org;
- Corpus queries: www.c-oral-brasil.org/db-com;
- C-ORAL-BRASIL I book download: https://www.dropbox.com/sh/s2n9w30dycnkauc/AAB1W_nIQtSVFdyPYeaGS2Fqa?dl=0

3. *NURC Digital*

Project NURC was a major corpus compilation initiative in Brazil, which started in 1969 with the aim of documenting and studying the spoken linguistic variety of five Brazilian state capital cities: Recife, São Paulo, Rio de Janeiro, Porto Alegre and Salvador.

The materials collected by project NURC have been used in a large number of academic papers, theses, dissertations and works of great importance, such as The Spoken Portuguese Grammar.

Project NURC Digital, which focused on the NURC data collected in Recife, had the following as its major goals: to digitize the entire data collection of the NURC Recife Project; to catalog and store all metadata in digital format; to propose a multilevel annotation system for the NURC Project data; to digitize the transcription data referring to the shared corpus of the NURC Recife Project; to archive all digitized data in international language databases and to make all NURC Recife material available on a dedicated website.

In order to achieve its goals, the NURC Digital Project followed the recommendations proposed by the Technical Committee of the International Association of Sound and Audiovisual Archives (IASA) and by the Open Archival Information Systems (OAIS). These recommendations were observed in all phases of the Project's development.

All the treated material were archived locally at the Federal University of Alagoas (UFAL) and at The Language Archive (TLA), based at the Max Planck Institute.

The NURC Digital audio files were transcribed in PRAAT (Boersma & Weenink 2015), then imported into ELAN along with the transcriptions in order for the sound-transcript allignment to be performed. The transcripts were parsed using the PALAVRAS parser (Bick 2000).

All materials from the NURC Recife Project are available in high quality digital format, annotated and revised. They encompass 346 recordings reaching 300 hs, 417 speakers, divided among 208 files. All multilevel annotations in TextGrid and eaf format are available for download.

The NURC Digital materials can be accessed and downloaded from the NURC Digital Portal at https://fale.ufal.br/projeto/nurcdigital/ as well as from The Language Archive at https://hdl.handle.net/1839/4C5B6AAD-97D8-4C53-846F-AB39FAD85F55.

4. *Corpus Brasileiro*

Corpus Brasileiro (The Brazilian Corpus) is the first mega corpus of Brazilian Portuguese. It was developed between May 2008 and April 2010, by Tony Berber Sardinha along with José Lopes Moreira Filho and Elaine Albert.

The corpus has had three versions. The first one, lauched in 2020, was hosted on a local server at PUC-SP, was tagged with TreeTagger and had a relational database format. Version 2 came out in 2012, had the same design as the previous version and was hosted in Linguateca and SketchEngine. The third version was released in 2015, with the same design, but tagged with PALAVRAS. The third version is a 10.5 GB download, with tagged and untagged versions, through gzipped files.

The Corpus Brasileiro design by mode portrays the following characterization: 8% spoken (83,055,313 words) and 92% written (1,005,163,599 words). The design by domain shows the following distribution: 53% academic, 24% news, 8% education, 8% politics, 4% encyclopedia, 1% technical, 1% literary fiction, 1% legal texts.

The corpus can be downloaded for free after a license form is filled out, through the link: https://form.jotform.com/51214092562953

More information about Corpus Brasileiro can be obtained from Berber Sardinha & Ferreira (2014) or through contacting the corpus manager, Telma São Bento Ferreira at telma.ferreira@corpuslg.org

5. *Linguateca Corpora*

Linguateca is an infrastructure project for Portuguese corpora and computational resources that is over twenty years old. It portrays both Brazilian and European Portuguese corpus resources, along with other Portuguese varieties.

One of its subprojects, AC/DC – "access and availability of corpora"contains materials developed by the Linguateca team as well as other corpus linguistics groups. All corpora in AC/DC are syntactically annotated and may also contain semantic annotation.

Some examples of Brazilian Portuguese corpora in AC/DC are: ReLi (Freitas *et al.* 2014), OBras (Santos *et al.* 2018) and DHBB (Higuchi *et al.* 2019).

ReLi is a book review corpus, containing 1,600 reviews and 133,000 words. It is a user-generated content corpus, comprised of book reviews posted on the internet. Its annotation portrays opinions about books and their polarities (positive or negative).

OBras is a Brazilian literature in the public domain corpus. It contains 272 texts, 43 authors and 7.2 million words. Obras was originally created to be the Brazilian counterpart of Vercial, a corpus of public domain literary works from Portugal. Its annotation encompasses places, people and human traits, literary genres (novel, short story, romance, etc.) and literary manifestation (realism, moderninsm, naturalism, etc.).

DHBB is comprised of Brazilian historical and biographical dictionaries and contains 7,685 entries and 9.8 million words. It belongs to the encyclopedic corpus genre and was designed to support Brazilian historical research and information extraction in the History domain. Its annotation portrays places, people, family relations, political parties and organizations.

The AC/DC corpora can be queried through the interface available at: https://www.linguateca.pt/acesso/info_acesso_English.php

6. *NILC Corpora*

The team at the Interinstitutional Center for Computational Linguistics (NILC) at USP-São Carlos have developed many Brazilian Portuguese corpora and computational resources. Three corpora will be presented below.

The first one is the PorSimplesSent corpus. This is a corpus comprised of aligned sentence pairs for the task of sentence readability assessment in Portuguese, as described in Leal *et al.* (2018). It is the first resource of this kind for the Portuguese language. The authors of the project made available four baselines for the corpus as well as an approach based on pairwise ranking to compare two versions of a sentence. Their model uses 17 lexical, syntactic and psycholinguistic features and identifies the readability level of sentence pairs with an accuracy of 74.2%. The corpus is publicly available at http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources.

The second corpus is the PropBank.Br (Duran & Aluísio 2011). This corpus project aimed at adding a layer of semantic role labels (SRL) to a treebank of Brazilian Portuguese. The first phase of such annotation provided a training corpus that is currently being used to develop SRL classifiers. The SRL annotation was added to the syntactic trees generated by the parser PALAVRAS (Bick 2000) in the Brazilian portion of Bosque, a manually revised subcorpus of Floresta Sinta(c)tica (Afonso *et al.* 2002). PropBank.Br 1.0 and 2.0 are available for download at: http://143.107.183.175:21380/portlex/index.php/en/downloadsingl

The third corpus is the CSTNews corpus (Cardoso *et al.* 2011), which is a collection of texts annotated according the CST (Cross-document Structure Theory) model. The CSTNews corpus contains 50 Brazilian Portuguese text collections. Each collection has approximately 3 documents on the same subject but from different sources. This corpus portrays a complex multilevel annotation system that encompasses: PoS and syntactic automatic annotation, wh-question/aspect annotation, text-summary sentence alignment, subtopics, noun and verb Wordnet senses, multidocument discourse annotation, single document discourse automatic summaries, and single and multidocument manual summaries. The corpus is available at https://sites.icmc.usp.br/taspardo/sucinto/cstnews.html

7. *Corpus do Português*

The Corpus do Português (Portuguese Corpus) encompasses a collection of corpora covering the following: Genre/Historical (45 million words), Web/Dialects (1 billion words), NOW (1.1 billion words). WordsAndPhrase (top 40,000 words).

The Genre/Historical corpus contains 45 million words of data from the 1200s-1900s, and it can be used to look at the history of Portuguese. For the 1900s, it is equally divided between spoken, fiction, newspaper, and academic texts, which means that it can be used to compare genres of Portuguese.

The Web/Dialects corpus contains about one billion words of data in web pages from four different Portuguese-speaking countries (Brazil, Portugal, Angola, Mozambique). This corpus allows a look at very recent Portuguese (the texts were collected 2013-14), and comparison among the different dialects.

In 2022, many new features were added to this corpus: 1) browsing and searching the top 40,000 lemmas in the corpus 2) detailed "word pages" with information on each of these 40,000 words, including definitions, synonyms, links to images and videos, frequency information (by genre and country), collocates, related topics, and concordance lines), 3) the ability to input and analyze entire texts, find keywords in these texts, and then see detailed information (cf. 2) for each word, as well as the ability to highlight phrases in a text and find related phrases in the corpus, and 4) extensive links to external resources in the frequency and conordance displays.

The NOW corpus is the newest addition to the Corpus do Português. It contains more than 1.1 billion words from four different Portuguese-speaking countries.

Finally, the WordandPhrase corpus allows searching and browsing through the top 40,000 words in Portuguese (based on frequency in the corpus). For each word, detailed information can be seen (all on one page) – definition, synonyms, frequency by genre, frequency by country, collocates (nearby words, which give great insight into meaning and usage), topics (co-occurring words on the same web pages), and 200 sample concordance lines (to see the patterns in which it occurs) – all with useful links from one word to another.

The Corpus do Português has a query interface that allows for several different types of searches. It is available at https://www.corpusdoportugues.org/x.asp

8. *Tycho Brahe*

The Tycho Brahe Parsed Corpus of Historical Portuguese (Galves 2019) is an electronic corpus of texts written in Portuguese by authors born between 1380 and 1978, encompassing both Portuguese and Brazilian authors.

At present, 88 texts (3,544,628 words) are available for research, with a linguistic annotation system in two stages: part-of-specch tagging (58 texts, a total of 2,280,819 words); and syntactic annotation (27 texts, a total of 1,234,323 words). The complete catalog of texts, along with author, title, number of words, as well as annotation status is available at http://www.tycho.iel.unicamp.br/~tycho/corpus/en/catalogo.html

The text preparation manual, along with the morphological and syntactic annotation manuals are available at http://www.tycho.iel.unicamp.br/~tycho/corpus/manual/index.html

The corpus page, through which all its specifications can be found, along with the download links is available at http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html

Acknowledgements

Mello and Raso thank FAPEMIG for the research grants that have enabled the development of the C-ORAL-BRASIL project. Berber Sardinha thanks FAPESP for the research grant that has enabled the development of Corpus Brasileiro.

Bibliography

Afonso, Susana & Bick, Eckhard & Haber, Renato & Santos, Diana. 2002. Floresta sintá(c)tica: a treebank for Portuguese. In *Proceedings of LREC-2002*. Available at: https://www.linguateca.pt/documentos/AfonsoetalLREC2002.pdf

Bick, Eckhard. 2000. *The parsing system PALAVRAS: automatic grammatical analysis of Portuguese in a constraint grammar framework*. Aarhus: Aarhus University Press.

Boersma, P. & Weenink, D. 2015. Praat: doing phonetics by computer. Available at: http://www.praat.org

Cardoso, Paula C.F. & Maziero, Erick G. & Castro Jorge, Maria Lucía R. & Seno, Eloize M.R. & Di Felippo, Ariani & Rino, Lucia H.M. & Nunes, Maria das Graças V. & Pardo, Thiago A.S.. 2011. CSTNews – A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In *Proceedings of the 3rd RST Brazilian Meeting*, 88-105. Available at: https://sites.icmc.usp.br/taspardo/RST2011-CardosoEtAl1.pdf

Cresti, Emanuela. 2000. *Corpus di Italiano parlato*. Firenze: Accademia della Crusca.

Duran, Magali Sanches & Aluísio, Sandra Maria. 2011. Propbank-Br: a Brazilian Portuguese corpus annotated with semantic role labels. In *Proceedings of the 8th Symposium in Information and Human Language Technology*, October 24-26, 2011, Cuiabá/MT, Brazil. Available at: https://aclanthology.org/W11-4519.pdf

Freitas, Cláudia & Motta, Eduardo & Milidiú, Ruy Luiz & César, Juliana. 2014. "Sparkling Vampire... lol! Annotating Opinions in a Book Review Corpus". In Aluísio, Sandra & Tagnin, Stella E.O. (eds.), *New Language Technologies and Linguistic Research: A Two-Way Road*, 128-146, Cambridge Scholars Publishing.

Galves, Charlotte. 2019. O corpus Tycho Brahe: um corpus sintaticamente anotado do português histórico. *Revista Binacional Brasil Argentina: Diálogo entre às Ciências*, 181-204.

Higuchi, Suemi & Santos, Diana & Freitas, Cláudia & Rademaker, Alexandre. 2019. Distant reading Brazilian history". In Navarreta, Costanza & Agirrezabal, Manex & Maegard, Bente (eds.), *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference (Copenhagen, Denmark, March 5-8, 2019)*, 190-200.

Janssen, Maarten & Freitas, Tiago. 2010. Spock – a spoken corpus client. In Oliveira Miguel Jr. (ed.), *Estudos de Corpora: da Teoria à Prática*, 11-126. Lisboa: Edições Colibri.

Leal, Sidney Evaldo & Duran, Magali Sanches & Aluísio, Sandra Maria. 2018. A nontrivial sentence corpus for the task of sentence readability assessment in Portuguese. In *Proceedings of the 27th International Conference on Computational Linguistics*, 401–413. Available at: https://aclanthology.org/C18-1034/

Plichta, Bartek. 2002. Best practices in the acquisition, processing, and analysis of acoustic signals. *University of Pennsylvania Working Papers in Linguistics* 8(3), Article 16. Available at:
https://repository.upenn.edu/pwpl/vol8/iss3/16

Raso, Tommaso & Mello, Heliana (eds.). 2012. *C-ORAL-BRASIL I. Corpus de referência do português brasileiro falado informal*. Belo Horizonte: Editora UFMG.

Santos, Diana & Freitas, Cláudia & Bick, Eckhard. 2018. OBras: a fully annotated and partially human-revised corpus of Brazilian literary works. In the public domain, *OpenCor*, Canela, RGS, Brasil, 24 de setembro de 2018.

Wittemburg, Peter & Brugman, Hennie & Russel, Albert & Klassmann, Alex & Sloetjes, Han. 2006. ELAN: a Professional Framework for Multimodality Research. In *Proceedings LREC 2006*. Available at: http://www.lrec-conf.org/proceedings/lrec2006/pdf/153_pdf.pdf