EXPERIENCE REPORT

BUILDING CAROLINA: METADATA FOR PROVENANCE AND TYPOLOGY IN A CORPUS OF CONTEMPORARY BRAZILIAN PORTUGUESE

Marcelo FINGER
Department of Computer Science - University of São Paulo (USP)
São Paulo, São Paulo, Brazil

Maria Clara Paixão de SOUSA
Department of Classical and Vernacular Letters - University of São Paulo (USP)
São Paulo, São Paulo, Brazil

Cristiane Namiuti
Department of Linguistic and Literary Studies - State University of Southwestern Bahia (UESB)
Vitória da Conquista, Bahia, Brazil

Vanessa Martins do MONTE
Department of Classical and Vernacular Letters - University of São Paulo (USP)
São Paulo, São Paulo, Brazil

Aline Silva COSTA
Coordination of Technical Courses in Information Technology - Federal Institute of Education, Science and Technology of Bahia (IFBA)
Vitória da Conquista, Bahia, Brazil

Felipe Ribas SERRAS
Department of Computer Science - University of São Paulo (USP)
São Paulo, São Paulo, Brazil

Mariana Lourenço STURZENEKER
Department of Classical and Vernacular Letters - University of São Paulo (USP)
São Paulo, São Paulo, Brazil

Miguel de Mello CARPI
Department of Computer Science - University of São Paulo (USP)
São Paulo, São Paulo, Brazil

Mayara Feliciano PALMA
Department of Classical and Vernacular Letters - University of São Paulo (USP)
São Paulo, São Paulo, Brazil

Gabriela Alves LACHI
Department of Oriental Languages - University of São Paulo (USP)
São Paulo, São Paulo, Brazil

EDITORS
- Raquel Freitag (UFS)
- Juliana Bertucci (UFTM)
- Márcia Vieira (UFRJ)

REVIEWERS
- Júlio Cesar Galdino (USP)
- Sérgio Serra (UFRJ)
- Mariana Costa (UFRJ)

ABOUT THE AUTHORS
- Marcelo Finger
  Conceptualization; Project Administration; Funding Acquisition; Resources; Supervision.
- Maria Clara Paixão de Sousa
  Conceptualization; Investigation; Methodology; Data Curation; Resources; Visualization; Writing - Original Draft.
- Cristiane NAMIUTI
  Conceptualization; Investigation; Methodology; Data Curation; Resources; Writing - Original Draft.
- Vanessa Martins do Monte
  Conceptualization; Investigation; Methodology; Data Curation; Resources; Writing - Original Draft.
- Aline Silva Costa
  Investigation; Data Curation; Formal Analysis; Software; Validation; Visualization; Writing - Original Draft.
- Felipe Ribas Serras
  Investigation; Data Curation; Formal Analysis; Software; Validation; Visualization; Writing - Original Draft.
- Mariana Lourenço Sturzeneker
  Investigation; Data Curation; Formal Analysis; Software; Validation; Visualization; Writing - Original Draft.
- Miguel de Mello Carpi
  Investigation; Data Curation; Formal Analysis; Software; Validation; Visualization.
- Mayara Feliciano Palma
  Investigation; Data Curation; Formal Analysis; Software; Validation; Visualization; Writing - Original Draft.
- Gabriela Alves Lachi
  Investigation; Visualization; Writing - Original Draft.

Received: 26/01/2025
Accepted: 17/07/2025
Published: 29/12/2025

HOW TO CITE
FINGER, M.; PAIXÃO DE SOUSA, M. C.; NAMIUTI, C.; MARTINS DO MONTE, V.; COSTA, A. S.; SERRAS, F. R.; STURZENEKER, M. L.; CARPI, M. de M.; PALMA, M. F.; LACHI, G. A. (2025). Building Carolina: Metadata for Provenance and Typology in a Corpus of Contemporary Brazilian Portuguese. *Cadernos de Linguística*, v. 6, n. 4, e812.

CHECK FOR UPDATES

ABSTRACT

This paper presents the challenges of building CAROLINA, a large open corpus of Brazilian Portuguese texts developed since 2020 using the Web as Corpus methodology enhanced with concerns about provenance and typology (*WaC-wiPT*). The corpus aims to serve both as a reliable source for research in Linguistics and as an important resource for Computer Science research on language models. Above all, this endeavor aims at removing Portuguese from the set of "low-resource languages". This paper details the construction methodology of CAROLINA, with special attention to the issue of describing provenance and typology according to international standards, while briefly describing its relationship with other existing corpora, its current state of development, and its future directions.

RESUMO

Este artigo apresenta os desafios da construção do CAROLINA, um grande corpus aberto de textos em português brasileiro em desenvolvimento desde 2020 que usa a metodologia 'Web as Corpus' aprimorada com preocupações de proveniência e tipologia (*WaC-wiPT*). O corpus pretende ser utilizado tanto como fonte confiável para pesquisas em Linguística quanto como importante recurso para pesquisas em Ciência da Computação sobre modelos de linguagem. Acima de tudo, este esforço visa retirar o português do conjunto das línguas de poucos recursos. Este artigo detalha a metodologia de construção do CAROLINA, com especial atenção ao problema da descrição de tipologia e proveniência segundo padrões internacionais; descrevemos também brevemente a sua relação com outros corpora existentes, seu estado atual de desenvolvimento e seus rumos futuros.

KEYWORDS

Brazilian Portuguese; Open Corpus; WaC; Typology; Provenance; WaC-wiPT.

PALAVRAS-CHAVE

Português do Brasil; Corpus Aberto; WaC; Tipologia; Proveniência; WaC-wiPT.

INTRODUCTION

CAROLINA is an open corpus for Linguistics and Artificial Intelligence, with a robust and unprecedented volume of texts of varied typology in Brazilian Portuguese. The current version, CAROLINA 1.3 Ada, is comprised of 802 million tokens, 2 million texts and more than 11 GB. All the texts were originally written in Brazilian Portuguese between 1970 and 2024, and are available for free, open access download at https://sites.usp.br/corpuscarolina. CAROLINA was conceived and is currently being developed by a multidisciplinary research team at the Digital Humanities Virtual Lab ('*Laboratório Virtual de Humanidades Digitais*', LaViHD) as part of the Natural Language Processing of Portuguese (NLP2) Project of the *Center for Artificial Intelligence* (C4AI) of the *University of São Paulo* (USP).

C4AI-USP endeavors to produce advanced research in Artificial Intelligence in Brazil, disseminate and debate the main results, train students and professionals, and transfer technology to society. The NLP2 Project, one of C4AI's challenges, seeks to develop systems that advance the state of the art of Natural Language Processing (i.e., NLP) to Brazilian Portuguese, targeting a new level of quality and performance compared to existing solutions. In this process, the Center aims to create opportunities for developing state-of-the-art language models and to distance Portuguese from the group of languages with "low NLP resources". With this aim, C4AI-USP, via the NLP2 Project, is currently building several Brazilian Portuguese corpora, including CORAA, *Corpus of Annotated Audios of spoken Portuguese*, and the *Portinari* annotated corpus of Portuguese. CAROLINA is C4AI's "mother ship" corpus and will incorporate CORAA's audio transcripts, *Portinari*'s raw unlabelled texts and other corpora in the future.

The corpus team, composed by computer science, linguistics and philology researchers, has worked to develop a methodology for building corpora that can be used in a variety of ways, complying with rigorous data control criteria in terms of its origin/provenance and typology, a fundamental requirement for not only computer science research but also linguistic research, among others. We aim at building a robust resource with state-of-the-art features both for research in the field of Artificial Intelligence and in the field of Linguistics, focusing on the importance of provenance and a rich typology of information as fundamental assets in modern data availability.

CAROLINA was named in honor of Carolina Michaëlis de Vasconcelos (1851-1925), a German philologist and linguist based in Portugal, and the first woman named a professor at the Faculty of Letters of the University of Lisbon, in 1911[1]. This tribute symbolizes the aims of our team: to

1 Carolina Michaëlis de Vasconcelos holds the distinction of being the first woman appointed as a university professor in June 1911, at the Faculty of Letters of Lisbon, although she never taught there. Preferring to remain in Porto, she requested and obtained a transfer to the Faculty of Letters of Coimbra, where she engaged in intensive teaching activity, leading the courses in Romance and Portuguese Philology (Sales, 2025).

advance knowledge on the Portuguese language and its history, and encourage scientific research made by women[2].

The aim of this paper is to present the challenges of building CAROLINA. The information presented here is very useful for creating future data sets, and it is also an invitation to use the corpus for developing researches in Linguistic area and others. For this purpose, section 1 presents the foundations on which this construction was based, section 2 shows how and why to develop a new methodology to build a giant corpus, highlighting the problems involved in the "Web as Corpus" idea, section 3 presents the current stage of the project, and section 4 concludes the paper with some final considerations and the indication of future steps.

FUNDAMENTALS

Since long before the emergence of the digital world, humanity has developed means and techniques to meet the need to organize, locate and retrieve documents and information. Knowledge of a text's source is directly related to the trustworthiness of its content. Thus, the provenance and typology of the documents figure in the range of essential information for the research in the Humanities and data reliability in Computer Science, especially for the construction of large collections of documents that serve to store knowledge in a recoverable, searchable and accessible way.

Linguistics research has largely benefited from digital technology, given that automation in the processing of large volumes of data strongly supports formulating hypotheses about grammars. In addition, the reliability of linguistic studies has been enhanced due to the development of scientific techniques and methods for annotation and data control from the sources of a natural language corpus. Paixão de Sousa (2014) underlies linguistic studies based on electronic corpora in the global approach of the text, in conceptual and technological terms, which is reflected in an interaction of different levels of analysis. Based on this global approach, CAROLINA has the potential to contribute to the development of research on Brazilian Portuguese, since it is being built aiming at reliability guarantees, assured by the provenance control provided by structured metadata.

According to Santos and Namiuti (2019), a scientific metadata control such as information about provenance and typology of the documents constitute essential information for the research in the Humanities and data reliability in Computer Science; in addition, it serves other areas, such as History

2 After naming the corpus, it came to our knowledge that Carolina's father, Dr. Gustav Michaëlis, was a mathematician, which brings an unexpected and felicitous relation to our work.

and Social Memory. To this end, the authors advocate the need for a structured metadata apparatus (AME - '*Aparato de metadados estruturados*') as a solution for reliability.

In Computer Science, NLP research has been dominated in the past years by a succession of *language models,* that is, a set of machine-learning architectures, mostly based on neural networks. Starting from sequence-to-sequence encoder-decoder (Kalchbrenner; Blunsom, 2013) configurations, it incorporated neural attention (Bahdanau *et al.*, 2014), leading to the attention-only Transformer architecture (Vaswani *et al.*, 2017). Different ways of assembling and training transformers have led to a multitude of very successful language models, such as BERT (Devlin *et al.*, 2019) and its derivatives (Liu *et al.*, 2019; Sanh *et al.*, 2019; Lan *et al.*, 2019) for text classification, GPT (Radford *et al.*, 2018) for text generation, and T5 (Raffel *et al.*, 2019) for machine translation. In such a rapidly changing environment, in which today's best model is condemned to short-term obsolescence, one must put forward a *language model training pipeline*, to be ready to generate "the next" proposed model. And the fundamental raw material for this production pipeline is a large, open and reliable general corpus such as ours.

CAROLINA was conceived within the Web as Corpus view (Baroni *et al.*, 2009), extended with provenance and typology information, which we call the *WaC-wiPT* view[3]. The Web as Corpus (WaC) view of corpus building (Fletcher, 2007) has been dominant in recent developments in linguistic resource building, but future applications may require a more cautious approach to data collecting. *Data provenance* refers to the process of tracing and recording the origins of data and, thus, its movement. It allows one to answer questions such as "**where did this piece of text come from?**", and "**is it a part or the whole of one document?**". Therefore, if a future application reveals that a corpus may carry some open or hidden biases, provenance is the mechanism that allows us to trace back the origins of the bias. It is also important to know if the data was transferred in total, or in part, and the size of the part. Data provenance provides the audit trail of the data and thus it is a source of reliability on data and on applications derived from it. Additionally, this work understands *typology* in a broad sense, as free from theoretical commitment as possible, and as a crucial methodological tool in the development of such a large collection of texts, organizing the search, selection and balancing of texts, as will be shown in the Methodology section.

Providing an open, large and diverse corpus for Brazilian Portuguese, with provenance and typology information has the potential of directly impacting research both on Linguistics and Computer Science. This is the intended goal of this work. We hope that provenance and typology information will be helpful to researchers. Control of information regarding digital documents produced or posted on the Web is necessary to meet the potential uses of a large collection of

3 A summarized version of the methodology developed can be found in Sturzeneker *et al.* (2022).

documents. This control also makes it possible to cater for a very wide range of research areas of interest, such as Social Memory and History, as well as Linguistics and Computing.

RELATED WORKS

Given the widespread availability of online content in the last decades, many researchers turned to the Web as their main source for corpus building. Examples of corpora that were built using the Web as a source are the *Terabyte* corpus (Clarke *et al.*, 2002) (53 billion tokens) and the *Web Tex*t Corpus (Liu; Curran, 2006) (10 billion tokens), both built using web-crawlers. The *Terabyte* Corpus targets the English language and is formed by HTML content obtained in mid-2001 from a set of URLs of the main sites of 2,392 universities and other educational organizations. The *Web Text* Corpus is also an English-language corpus composed of a collection of texts on various subjects. Unlike most corpora created for NLP use, this corpus employs a linguistic search process instead of the traditional use of web search engines, which are based on scores. The general objective of the corpus was to measure the accuracy of NLP-learning software using it in comparison to training using other corpora (Liu; Curran, 2006).

The *TenTen* Corpus Family is an initiative by Sketch Engine[4] for building Web corpora for all major languages in the world (Jakubíček *et al.*, 2013). The *TenTen* corpora was also created using web-crawling techniques and presenting texts validated under exclusively linguistic criteria, using specialized technology for this. The very name of the corpora (*TenTen*, 10^{10}) indicates the large size of each corpus that composes it, starting from the minimum size of 10 billion words for each language[5].

In this context, several large corpora were built adopting the WaCky (*Web-As-Corpus Kool Yinitiative*) methodology, following the emergence of the first WaCky corpora: the *ukWaC*, *deWaC*, *itWaC* (Bernardini *et al.*, 2006; Baroni *et al.*, 2009)**,** and the *frWaC* (Ferraresi *et al.*, 2010), which targeted English, German, Italian, and French respectively and include more than 1 billion words each. This methodology comprises four steps which are: identification of different sets of seed URLs, post-crawl cleaning, removal of duplicate content, and annotation. One of the corpora built following this framework is the *Brazilian Portuguese Web as Corpus* (*brWaC*), of great relevance

4 https://www.sketchengine.eu/

5 The TenTen Corpus Family is available for consultation in more than 40 languages, including a corpus of 4 billion tokens for Portuguese (ptTenTen), which includes the European and Brazilian variants. However, the corpora of all these languages are not openly available, being accessible only from the Sketch Engine platform (Jakubíček *et al.*, 2013; Wagner Filho *et al.*, 2018).

as it was already considered the "biggest Brazilian Portuguese corpus available" during its construction (Boos *et al.*, 2014)[6].

The use of the web as a data source for corpus construction, made possible by the WAC methodology, called for a progressive redefinition of the traditional concept of a corpus – see the differences in corpus definitions between Sardinha (2000) and Kilgarriff and Grefenstette (2003) – while providing linguistic research with click-through access to large-scale collections of real-world language samples. This, in turn, facilitated the detection of highly uncommon linguistic patterns at levels of representativeness previously unattainable (Kilgarriff; Grefenstette, 2003).

These advances, however, have not been without challenges. Critics of the approach highlight the high incidence of typographical and extraction errors, the differences in distribution between web-based language and that used in other contexts, and the challenges of defining and capturing the typological distribution of virtual language (Kilgarriff; Grefenstette, 2003).

It is also worth noting that, in its early days, the web was perceived as an easily accessible copyright-free repository of language, in contrast to the professionally edited sources previously available for corpus construction (Kilgarriff; Grefenstette, 2003). At the present time, however, debates over the right to build datasets from web content, particularly at a time when such resources are used to train large proprietary language models, are at the heart of ethical and legal controversies. Carolina, with its collection policy grounded on license compliance and provenance, has been developed with the explicit aim of addressing this issue.

Regarding other existing Portuguese-language corpora, the virtual organization *Linguateca* (Santos, 2000) stands out as a center for resources focused on the computational processing of this language. Its objective was to contribute to the development of new computational and linguistic resources, facilitating the access of new researchers to existing tools. Of the corpora available at *Linguateca* that specifically target Brazilian Portuguese, the ones that stand out as the most significant in size are: *Brazilian Corpus* (Sardinha; Filho; Alambert, 2010), *Lácio-Web* (Aluísio *et al.*, 2003), and *Corpus do Português*, subcorpora *NOW* and *Web/Dialects* (Davies; Ferreira, 2016, 2018).

6 Published in 2017, it contains 2.68 billion tokens that were crawled from the Web in 24 hours, by initially employing queries to a search engine with random pairs of content words, according to the description of its development in Wagner Filho *et al.* (2018). Its importance for advances in Brazilian Portuguese research in multiple areas is illustrated by its employment in NLP model training, as substantiated in Souza, Nogueira and Lotufo (2020).

The *Brazilian Corpus* has approximately one billion words, syntactically annotated with the parser *PALAVRAS* (Bick, 2000).[7] *Lácio-Web* was developed by USP as a project whose objective is to make fully and freely available its linguistic and computational tools as well as several corpora of contemporary Brazilian Portuguese. This set of corpora prioritizes whole-content texts and a variety of genres, text typologies, domains of knowledge, and means of distribution (Pinheiro; Aluísio, 2003; Aluísio *et al.*, 2004).[8] *Corpus do Português* was developed by Brigham Young University and Georgetown University. The *NOW* (*News on the Web*) subcorpus has approximately 1.1 billion words of four different Portuguese varieties (Angola, Brazil, Mozambique, and Portugal), gathered from daily searches of articles in magazines and newspapers through Google News between 2012 and 2019. It is not possible to easily retrieve the source and copyright information of the texts, nor to know how much of the data refers specifically to Brazilian Portuguese. The subcorpus *Web/Dialects*, in turn, has approximately one billion words of the same four Portuguese varieties, of which 656 million words are in Brazilian Portuguese, mainly extracted from Blog-type sites (Davies; Ferreira, 2016)[9].

1. BUILDING A METHODOLOGY

The construction of a billion-token corpus requires a considerable amount of preparation and coordination. First of all, we had to define the metadata scheme, which was adjusted after the initial surveys and tests. The goals of the corpus must remain clear at all times, and a mechanism for tracing sources, completion levels and data balancing must be followed diligently. Such an endeavor required three important stages: a detailed analysis of existing resources, the development of a methodological framework to adhere to our goals, and the developing of techniques for post-processing. The main methodological decisions in this process are described in **1.1** and **1.2** below, with

7 It was developed by the Applied Linguistics and Language Studies Program of the Pontifical Catholic University of São Paulo (LAEL/PUC-SP) and its version 6.0 of February 2, 2020 is available for online searches at Linguateca. The full corpus can be downloaded upon approval of a requisition form, as long as the user agrees not to distribute or use it for profit purposes. Although the primary sources of the texts are not explicit, there are several works on the construction of the Brazilian Corpus that describe the set of typologies and textual genres that compose it (Sardinha *et al.*, 2010; Vianna; de Oliveira, 2010; de Oliveira; Dias, 2009; de Brito *et al.*, 2007).

8 The Lácio-Web project comprises six corpora, four of which are currently available online (Lácio-Ref, Mac-Morpho, Par-C, and Comp-C) and whose content is described on its website: http://143.107.183.175:22180/lacioweb/descricao.html.

9 In both corpora, the texts were processed after the download for boilerplate removal, duplicates detection, lemmatization and tagging. The complete Web/Dialects corpus is available for purchase, with different selling prices depending on the final license chosen by the user (academic or commercial) at https://www.corpusdata.org/purchase.asp. The NOW Corpus can be accessed free of charge for online searches on its website (https://www.corpusdoportugues.org/now/), but it cannot be downloaded in full.

special attention to aspects related to Provenance and Typology, and the processing stages are presented briefly in **1.3** and **1.4** further on.

1.1. THE ISSUE OF PROVENANCE

Initially, significant effort was made to analyze the pre-existing resources for natural language processing in Brazilian Portuguese, with the aim of supporting the development of our methodology and exploring the possibility of incorporating some of these resources into Carolina. That enabled us to assess the benefits and drawbacks of their methodologies, as well as which niches of contemporary text were already corpus-indexed, and which were still fertile sources for us. In doing so, we decided on a web-based corpus, but against the adoption of the WaCky framework.

Despite the WaCky method claiming the facilitation of an automatic balancing of content without bias, and the brWac achieving the reduction of limited-relevance and duplicate content in comparison to other WaCKy corpora (Wagner Filho *et al*., 2018), the methodology presents some drawbacks. As their creators acknowledge, automated methods offer limited control over the content included in the final corpus, thus necessitating post-hoc investigation (Baroni *et al.*, 2009). For example, the *brWac* researchers only provide the annotated categories of the 100 most frequent websites (Wagner Filho *et al*., 2018), and unlike the other WaCky corpus mentioned in the previous section, the list of bigrams, seeds and total URLs used for the *brWaC* construction are not easily accessible.

These challenges for content quality and provenance tracking, as well as on rights of use, are central issues in CAROLINA'S goals, and our methodology was developed around avoiding such problems. In line with Paixão de Sousa (2014) and Santos and Namiuti (2019), a scientific control of the memory processes of building a corpus, the memory of texts, and the definition of the set of essential metadata to control provenance and guarantee the documents' reliability figure in the range of essential information for the research in Humanities and in Computer Science. Furthermore, as we intend to openly distribute the contents of the corpus online, under terms akin to Apache license and similarly permissive ones, assuring data provenance beforehand is crucial to determine the original terms of use of content.

For instance, Davies and Ferreira (2018) recognize that the texts used in the *Brazilian Corpus* may be copyrighted and, for this reason, they rely on the American Fair Use Law, which states that texts under copyrights can be used freely as long as their format is remodeled and that there is no foreseen impact on their potential market by their legal holders (Stim, 2016a; 2016b). Thus, to avoid copyright-related problems in the *Brazilian Corpus,* for every 200 words of text, 10 words were replaced by "@", totaling the exclusion of 5% of the original text. The authors claim that, as the words

were removed regardless of context, all words would be affected equally, so the frequency and usage counts would not be affected and the corpus would still be suitable for linguistic studies[10].

However, depending on local legislation and on the purpose of the collected material, corpora might still violate the law, even when publishing only fragmented or highly processed versions of texts that include copyrighted content. In addition, when crawling based on seed URLs and search engines, there is no control over the copyright nature of texts. According to Cardoso (2007), Brazilian law acknowledges copyright establishment at the exact moment an intellectual work is created, without the need for further legal requests or paperwork. Therefore, being published and openly accessible online is no waiver of copyright limitations. For this reason, while building CAROLINA we avoided collecting random samples from the web to ensure both the openness of the information crawled and compliance with Brazil's recently enforced personal data protection law, LGPD: "*Lei Geral de Proteção de Dados*"[11].

As for the possibility of incorporating pre-existing corpora to CAROLINA, there are some obstacles to be considered. Firstly, many corpora listed at *Linguateca* were discarded from our list for not fitting into our date range or for having content that may have reproduction and distribution restrictions due to possible copyright attributed to the texts or to the corpus itself. Many corpora that took into consideration the copyright limitations chose to work solely with fragments or excerpts of the original texts, choosing greater ease of access to the text at the expense of the completeness of its content, as is the case of *Corpus do Português*. In the construction of the corpus, we prioritize the use of integral or minimally modified texts, as we understand that a fragmentary nature of the content can be detrimental to inter-phrase or inter-text associations in both linguistic studies and software development for Natural Language Processing, amongst which the most recent alternatives, such as Attention-based Algorithms, require the processing of the text sentences in its completeness (Devlin *et al.*, 2019). Another obstacle for corpus incorporation is that a large part of the datasets and smaller corpora gathered at *Linguateca* uses the European variety of the Portuguese language, or more than one variety. This constitutes a limitation to their incorporation, considering that CAROLINA intends to be an open corpus of contemporary Brazilian Portuguese.

Thus, we concluded that it would be more productive not to include the large existing corpora of Brazilian Portuguese in CAROLINA, but rather use them as theoretical guidance and control parameters for the development of a new methodology based on provenance, typology and free distribution of the texts: the WaC-wiPT.

However, since CAROLINA is C4AI's "mother-ship" corpus, we may incorporate additional smaller corpora in the future, expanding upon those already included in our broader typology, Datasets and

10 https://www.corpusdata.org/limitations.asp

11 Available in Portuguese at http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm.

Other Corpora. We are interested mainly in those corpora whose content is unique or not easily independently recoverable, such as corpora of transcribed spontaneous speech, like the corpora developed by Project TaRSila[12], at C4AI, already described in Santos *et al.* (2022). We believe that these unique-content corpora will be important sources to guarantee a greater representation of dialectal and typological varieties to the corpus.

1.2. DEFINING TYPOLOGIES

Having determined the objectives and philosophy for the construction of CAROLINA, we aimed to build it with reliability guarantees ensured by provenance control through structured metadata. To achieve this, we focused on conducting surveys by broad typologies, which we defined as a way to group related web domains with similar content. After defining a typology methodology, we started the downloading step, followed by a preprocessing phase and, finally, we proposed the categorization of metadata headers and the metadata scheme.

The surveys started by a broad typology (Figure 1), divided in seven types, as detailed below. The seven broad types first defined were categories that allowed us to group all the domains researched up to that point: judicial branch of government; legislative branch of government; datasets and other corpora; public domain works; wikis; university domains; and journalistic texts. As we chose the sources to be surveyed for the broad typology, we gave priority to those with open data and a large volume of files, since the process of requisition of rights of use would only take place in later stages of the project. Therefore, the sources that had copyright-protected data (for instance, the journalistic texts) were not prioritized in the first moment (Crespo *et al.*, 2022).

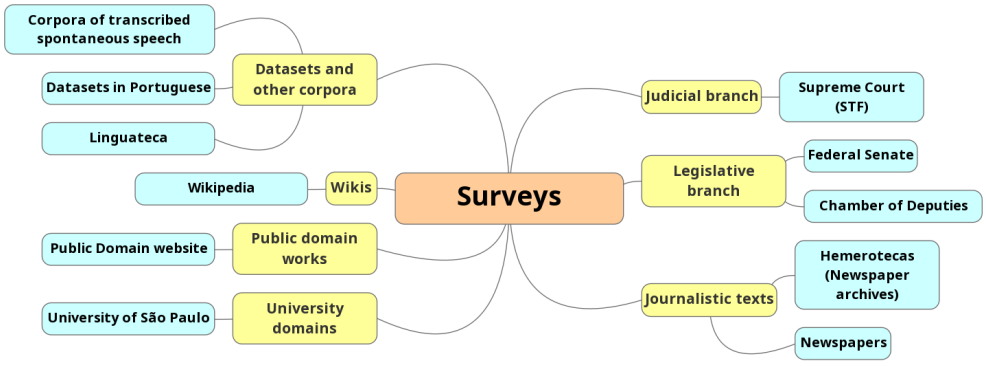

Figure 1. Surveys by broad typology.

12 https://sites.google.com/view/tarsila-c4ai

The surveys consist of an in-depth research of each broad type chosen for the construction of the corpus and investigation of the web domains that comprise them. Thus, we surveyed information about the license of the texts and the basic directory structure of the investigated sites, as well as authorship, date, and other information that we deemed relevant for each broad type. All this collected data involved a great importance to the download process and it has been essential for the insertion of the predefined metadata and their revision. Therefore, the surveys are continuously ongoing, as research must be conducted or supplemented for each new web domain we wish to incorporate into Carolina.

In addition, given that throughout the surveying process we came across various types – often within a single web domain –, we defined a narrow typology, formed by subdivisions of the broad typology that take into account the structural similarity of the extracted files. Therefore, we created the distinction between broad and narrow typology: the former is an initial web domain grouping by similar content, and the latter, a more detailed label for the types of texts found in each section or file of a surveyed web domain. Narrow typology was also included in most surveys and is a metadata category.

1.3. THE DOWNLOADING AND PREPROCESSING STAGES

After the initial survey by broad typology, the downloading and preprocessing stage begins to take place. It is important to note that those procedures, which could be called the 'final' stage of the corpus construction, rely heavily on our principles of Provenance and Typology. As discussed in section 2.1, text provenance is the baseline criterion for a text to be selected for the corpus; and as shown in section 2.2, the broad typologies of the texts are the guidelines over which the building process begins. The downloading and preprocessing stages described here were the basis for the production of versions CAROLINA 1.0 Ada (2021), CAROLINA 1.1 Ada (2022), CAROLINA 1.2 Ada (2023) and CAROLINA 1.3 Ada (2024). CAROLINA 2.0 Bea is being prepared for publication in 2025.

The files were mostly obtained through *Wget*, the chosen software for this process. As the Raw Corpus[13] (which precedes text preprocessing) aims at safekeeping the entirety of the selected web domains, thus avoiding any future problems in case they are partially or completely removed from the Internet, the mirror command was used. This command crawls entire web pages, with infinite recursivity inside a web domain, creating by default a mirror of its directory structure, complying with our intention of archiving a copy of most of the sources used in the corpus.

13 *Raw Corpus* ('*Corpus Cru*') is a concept created to describe a product derived from the Lapelinc method for corpora building (Namiuti; Santos, 2021), which consists of an unpublicized version of the corpus that holds more information about itself, serving as a mirror of the original sources of the corpus. This notion has been adapted to the methodology used to build CAROLINA.

With the detailed inspection of each type in the broad typology, the process of downloading the files was facilitated. Accordingly, in some cases, pages whose content was irrelevant or out of the proposed frame were ignored, such as the public domain works[14] published prior to 1970. In those cases, the files were assessed one by one and downloaded with *Wget*, by means of feeding it a URL list in a TXT file.

The filtering of texts was included in the process of building the corpus version and is based on surveys of each type within the broad typology. Care was taken to exclude anything outside the proposed time span (1970–present) and pages with little or no textual content. Therefore, as these previous inspections enable a closer understanding of the structure of the surveyed websites, the desired sections will be easily tracked and selected for preprocessing among the downloaded files. In the preprocessing stage, we extract the text from the Raw Corpus and, after that, the Metadata insertion process takes place.

That methodology was also relevant when the mirror command did not retrieve all the targeted files of a website. As the initial survey allowed for the learning of which pages or directories were desired for the corpus, other download methods had to be employed in the cases where they were not automatically crawled by the mirror command. This difficulty was present especially in the Brazilian Federal Government public websites, which required alternatives to obtain their content, and many resources were used for that. For different sections of the Brazilian Supreme Court (*STF*) website[15], for example, we built tools to generate URLs based on file naming patterns, to extract URLs and save pages with an HTML parser, and to access and click links recursively, using the Python[16] library *BeautifulSoup* and the *Selenium WebDriver*. In addition to that, a large volume of judiciary documents was kindly provided by Jackson José de Souza, crawled with a tool he developed using *Scrapy*.[17]

1.4. DEFINING METADATA

The stage following preprocessing is metadata insertion. The conception and development of appropriate Metadata categories have been core tasks in building CAROLINA. The identification of basic metadata for the objectives of the corpus was guided by the classification of information into two broad categories. The first category groups objective information contained in the source document of the text, not having been generated from any type of analysis. Following Santos and

14 For the time being, all of the files of this broad typology were extracted from http://www.dominiopublico.gov.br/.

15 http://portal.stf.jus.br/

16 All the tools were developed using Python 3.

17 The tool he developed for his Master's degree at the University of São Paulo is available at https://github.com/mayara-melo/analise-juridica.

Namiuti (2019), we name this category "Dossier". The second category, which we name "Carolina", includes processing information and information generated from the analysis of the text contained in the source document. From these two categories, eight information groups were identified: Primary Identification, Authorship, Dating, Location, Size, Acquisition, Licenses and Typology.

Table 1 lists each piece of metadata identified as necessary for the corpus text header. The first column shows the information category ("Dossier" or "Carolina"); the second column identifies the information group within each category; the third column specifies the item of metadata within each group. The last column specifies the cardinality of each metadata, determining the minimum and maximum or the exact number of occurrences of that metadata for each file: the minimum cardinality of "zero" indicates that the item of metadata is optional, while the maximum cardinality of "one", or "one or more" indicates that it is mandatory.

Category	Group	Item of Metadata	Cardinality
Carolina	Primary Identification	Name of the file created in the corpus	1
		Corpus description	1
	Authorship	Credits for work with corpus file (download, extraction and metadata insertion)	1+
		Authority responsible for the file in the corpus	1
	Date	Download date of the source file	1
		Extraction date	1
	Licenses	License type of the file created in the corpus	1
		Access conditions for the file created in the corpus	1
	Extent	Number of tokens in the text from the source document	1
	Typology	Carolina typology	1
Dossier	Primary Identification	File name in the corpus	1
		Title of the source document	0
	Authorship	Source document author	0+
		Source document translator	0+
		Authority responsible for the source document	1+
		Sponsor (Institution responsible for the source document)	1+
		Publisher	0
	Date	Date or period (start and end) of the source document	0

	Licenses	License of source document (Public domain, Commons, etc.)	1
		License URL	1
		Access conditions of source document (free or restricted)	1
	Localization	Source document access URL	1
		Regional origin of source document	0
	Acquisition	File format of the source document (pdf, html, etc.)	1
		Constitution (integral, fragmented, etc.)	0
		Nature of acquisition (native digital, scanned printout, OCR)	0
		Part (part of the collection the document represents, if app.)	0
		Collection (of which the document is part of, if app.)	0
	Extent	File size in bytes of the source document	1
		Number of pages of the source document	0
		Number of tokens of the source document	1
		Originality of the source document measured with Onion18	1
	Typology	Document type declared in the source document	0
		Linguistic variety (regional) indicated in the source document	0
		Written, spoken or mixed text (transcribed)	1
		Domain of use	1
		Preparedness (spontaneous, monitored, mixed or unknown)	0
		User-generated	0

Table 1. Identification of metadata for CAROLINA (1.3 – Ada Version).

The texts in CAROLINA are represented as TEI Documents, encoded in XML in accordance with the specifications "TEI P5: TEI Guidelines for Electronic Text Encoding and Interchange", developed and maintained by the Text Encoding Initiative Consortium (TEI Consortium, 2024). A single XML file encodes several texts of the corpus, with a hierarchy of elements that can be validated against a schema derived from the TEI Guidelines, aiming to ensure greater interoperability.

18 The "Originality" metadata represents the percentage of non-duplicated content of a text in relation to the whole corpus. It is measured using the Onion corpus tool (https://corpus.tools/wiki/Onion).

Each text included in the corpus contains a <TEI> element, which includes the descendant node <teiHeader>, mandatory in a TEI-conformant document. The metadata items listed in Table 2 are encoded in <teiHeader> element for each text. Figure 2 presents the general hierarchical structure of the XML header structure of each individual text. Carolina's header of text was structured based on the AnnoTEI Schema, proposed by Costa (2024), which recommends encoding each item of metadata classified in the "Dossier" category within the <sourceDesc> (Source Description) element. Given the importance of the origin of the texts for the Carolina project, even though they are native digital documents, the data referring to the source document or file are encoded in sourceDesc. The <fileDesc> element (File Description) constitutes a mandatory element into the <teiHeader> and is designed for encoding file description. Since the <sourceDesc> element can contain <fileDesc>, the AnnoTEI Schema defines that this element includes a complete bibliographic description of the source document, while the remaining metadata items are inserted into other child elements of <fileDesc> that are not nested in <sourceDesc>, which includes information about the distribution and working with the corpus XML file.

Building on this, the final schema for the texts in the corpus was defined observing the specificity of the project. CAROLINA'S <teiHeader> also contains two elements defined as optional by TEI Guidelines: <encodingDesc> (Encoding Description) and <profileDesc> (text-profile Description). The <encodingDesc> element includes information about the encoding of the text. Finally, profileDesc contains the text classification according to the typology established by the project team. The decisions of which elements to use and their location were based on the objectives of the corpus, creating a schema in accordance with the "corpus" customization provided by the TEI.

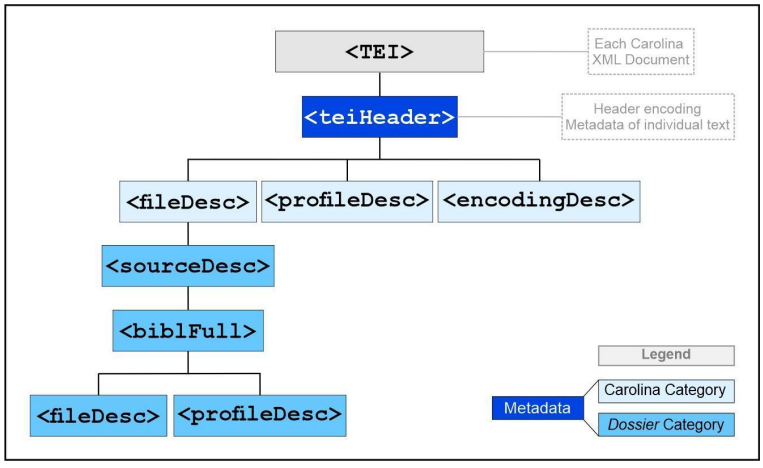

Figure 2. General hierarchical structure of the CAROLINA XML file.

Since TEI P5 is a modular and flexible system, whose infrastructure enables users to create a specified encoding schema appropriate to their needs without compromising data interoperability, a customized schema was defined for the CAROLINA CORPUS. The final schema meets the

conformance requirements outlined by the TEI standard, ensuring that documents validated by it are "TEI-Conformant". Therefore, the customized schema follows the TEI Abstract Model and is generated from an ODD (One Document Does It All) file, as recommended by the guidelines. To achieve greater interoperability, the customization is a subset of the "TEI-All" schema, which makes it "clean modification", according to the guidelines (TEI Consortium, 2024). This conformance ensures that the metadata encoded in the <teiHeader> can be consistently mapped onto widely adopted standards such as Dublin Core and, when required, onto CMDI, the standard adopted by the European CLARIN infrastructure.

The TEI header was designed to ensure interoperability with bibliographic and linguistic metadata standards. Since the AnnoTEI schema does not represent an extension of TEI, but rather what the guidelines classify as a *clean modification*, the <teiHeader> used in Carolina is fully compatible with international metadata standards and can be entirely mapped to Dublin Core. Furthermore, by selecting an appropriate profile, the header can also be converted to CMDI, the metadata standard adopted by the European CLARIN infrastructure for language resources. The Carolina corpus is already available in Portulan CLARIN, the Portuguese node of this infrastructure, thereby ensuring its interoperability with international metadata standards.

2. CURRENT STATE

Since 2022, four versions of CAROLINA ADA have been published, each one with few updates or corrections in relation to the previous version. Table 2 below shows the schedule and size of each version, more information about all of them you can find on the corpus' webpage (https://sites.usp.br/corpuscarolina).

Date	Carolina Version	Size (GB)	Number of extracted texts	Number of tokens[19]
2022, Mar	1.0 - Ada	39.23	1,745,234	653,346,569
2022, Apr	1.1 - Ada	7.2	1,745,187	653,322,577
2023, Mar	1.2 - Ada	11.36	2,107,045	823,198,893
2024, Oct	1.3 - Ada	11.16	2,076,205	802,146,069

Table 2. Published CAROLINA Versions.

19 The token count was achieved with the wc -c linux command, which counts "whitespace-separated tokens": https://www.gnu.org/software/coreutils/manual/html_node/wc-invocation.html#wc-invocation. It means that these sums will significantly decrease after the preprocessing phase. As well as the number of files, yet we expect it to be in a lower range than the 95% of discarded documents crawled by the brWaC (Wagner Filho *et al.*, 2018), for instance. This same observation is valid for Table 3.

The current version of the corpus (Carolina 1.3 Ada), published in October 2024, is organized by the types in the broad typology established up to the present, plus an additional Social Media typology, and it shows the following numbers (Table 3).

Broad typology	Size (GB)	Number of extracted texts	Number of words
Datasets and other corpora	4.3	1,074,032	196,524,339
Judicial branch	1.5	40,398	196,228,167
Legislative branch	0.025	13	3,162,474
Public domain works	0.005	26	601,465
Social Media	0.017	3,294	1,231,881
University domains	0.011	941	1,078,967
Wikis	5.3	957,501	403,318,776
Journalistic texts	0	0	0
Current total	**11.16**	**2,076,205**	**802,146,069**

Table 3. XML CAROLINA Corpus (Version 1.3 Ada) in numbers.

The information presented in Table 3 concerns the XML corpus, which represents the final stage of CAROLINA 1.3 Ada version, containing texts extracted from open source web domains, balanced data, and their respective metadata encoded. The texts in the XML version have already been preprocessed, filtered, and deduplicated. Besides, it is valuable to mention that some websites or even entire broad types (such as the journalistic texts), which require the explicit authorization of their copyright owners to be made available (and, therefore, are still in the course of requisition of rights of use), are not being accounted for in the numbers.

The work on building Carolina is constantly ongoing and a new version with its own search interface is being prepared.

3. CONCLUSION AND FUTURE STEPS

CAROLINA has an important distinguishing feature: it is conceived with an original methodology developed by the LaViHD-C4AI team, which we call *WaC-wiPT* (*Web as Corpus with Provenance and Typology information*). We consider provenance to be a crucial aspect to strive for in web-based corpora, alongside typology and balance management. Apart from facilitating copyright compliance and typology labeling, it allows one to answer questions about the origin of texts and increases the scope of uses for the corpus.

As shown in our state-of-the-art review (a non-exhaustive list of openly available Brazilian Portuguese corpora and other relevant web-based corpora), many recent corpora were built adopting the WaCky methodology. Because this methodology does not envision provenance as we defend here for CAROLINA, and because most guidelines from other corpora emphasize only "corpus balance", for which typology serves just as one criterion, most of these corpora were not incorporated into the corpus; instead, they had an important role in the conception of our own methodology.

Therefore, in line with the provenance proposition, the LaViHD team at C4AI, as part of its NLP2 challenge, has built a large corpus with a robust and unprecedented volume of texts of various typologies in Brazilian Portuguese. In this paper, we presented the current state and the next steps of the corpus construction, defending the importance of provenance and of a detailed typology scheme as fundamental assets in modern data availability. As related products developed during the construction of CAROLINA, we presented the WaC-wiPT methodology, based on provenance and typology, aiming to make as much data as possible openly available online (in its "beta version"). This also includes the building of metadata to describe provenance and typology, forming the first version of CAROLINA'S header scheme, using "TEI P5" in accordance with the reuse principle. Additionally, the Raw Corpus was created, currently totaling 1,779 GB and 124,084,164,722 tokens.

As CAROLINA approaches its fifth anniversary, we hope to be very aware of its limitations as well as its progress. In this regard, one of the main challenges at the current phase is balancing the corpus in terms of typologies; in particular, as we mentioned before, the sources of texts that had copyright-protected data (for instance, the journalistic texts) were not prioritized in this first moment. We are aware that this limitation must be overcome, and this is part of our goals for the next versions. Interestingly, we observe that this problem stems from our principle of guaranteed Provenance; but rather than compromise on this fundamental aspect, we opted to wait some time until we can obtain the correct licenses that would allow us to offer quality, whole texts independent of copyright liability.

Finally, another important challenge in the current phase of the development is the availability of a more user-friendly interface, in particular bearing in mind the users outside the realm of Computer Science. In its current state, CAROLINA is fully available through the main website, leading to dedicated

platforms which allow bulk download[20]; in the near future, we will make it available on a searchable interface which will complement the possibility of downloading the whole corpus.

ACKNOWLEDGMENTS

This work was carried out at the Center for Artificial Intelligence (C4AI-USP), with support by the São Paulo Research Foundation (FAPESP grant #2019/07665-4) and by the IBM Corporation. This work was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior -- Brasil (CAPES) -- Finance Code 001 and also by the Ministry of Science, Technology and Innovation, with resources of Law N. 8.248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC 13, DOU 01245.010222/2022-44. Marcelo Finger received partial support from FAPESP (\#2023/00488-5, \#2022/11254-2) and CNPq 302963/2022-7 (PQ), Cristiane Namiuti received partial support from Bahia Research Foundation (FAPESB 0007/2016, 0014/2016), Maria Clara Paixão de Sousa and Vanessa Martins do Monte received partial support from the São Paulo Research Foundation (FAPESP grant #2021/15133-2), Felipe R. Serras was supported by the IBM Corporation in a grant managed by FUSP under number 3541 and in a PPI-SOFTEX grant managed by FUSP under number 3970, Mariana L. Sturzeneker received support from the São Paulo Research Foundation (FAPESP grant #2024/13270-0) and Gabriela Alves Lachi received support from the Unified Scholarship Program of the University of São Paulo (PUB-USP), project 2024-5789.

We would like to thank the researchers who were involved in the earlier phases of Carolina but are no longer part of the project today: Maria Clara Ramos Morales Crespo, Maria Lina de Souza Jeannine Rocha, Guilherme Lamartine de Mello, Raquel de Paula Guets, Renata Morais Mesquita, Mariana Marques da Silva and Patrícia Brasil. Their contribution was essential to getting us to where we are now.

20 Portulan Clarin (https://portulanclarin.net/repository/browse/carolina-general-corpus-of-contemporary-brazilian-portuguese-with-provenance-and-typology-information/f3751b34e36611ecaa5802420a870112f00a37650c304dbda703d85e14a2e945/) and Hugging Face (https://huggingface.co/datasets/carolina-c4ai/corpus-carolina)- cf. full list of all versions available for download at https://sites.usp.br/corpuscarolina/corpus.

ADDITIONAL INFORMATION

CONFLICT OF INTEREST

The authors declare no conflict of interests.

STATEMENT OF DATA AVAILABILITY

This research is conducted as an Open Access Project.

FUNDING SOURCES

- IBM Corporation.
- Brazilian Ministry of Science, Technology and Innovation.
- São Paulo Research Foundation (FAPESP), grants 2019/07665-4, 2021/15133-2, 2022/11254-2, 2023/00488-5, 2024/13270-0.
- University of São Paulo Support Foundation (FUSP)
- USP Unified Scholarship Program (PUB)
- Coordination for the Improvement of Higher Education Personnel (CAPES), Finance Code 001.
- National Council for Scientific and Technological Development (CNPq), 302963/2022-7 (PQ).
- Bahia Research Foundation (FAPESB), grants 0007/2016, 0014/2016.

REVIEW AND AUTHORS' REPLY

Review: https://doi.org/10.25189/2675-4916.2025.V6.N4.ID812.R
Author's Reply: https://doi.org/10.25189/2675-4916.2025.V6.N4.ID812.A

REFERENCES

BICK, E. *The parsing system palavras*: Automatic grammatical analysis of Portuguese in a constraint grammar framework. Aarhus: Aarhus Universitetsforlag, 2000.

BOOS, R.; PRESTES, K.; VILLAVICENCIO, A.; PADRÓ, M. brWaC: A WaCky Corpus for Brazilian Portuguese. *In*: BAPTISTA, J.; MAMEDE, N.; CANDEIAS, S.; PARABONI, I.; PARDO, T.A.S.; VOLPE NUNES, M.G. (Eds). *Computational Processing of the Portuguese Language. PROPOR 2014, Lecture Notes in Computer Science, vol 8775*. Cham: Springer, 2014. p. 201-206. https://doi.org/10.1007/978-3-319-09761-9_22.

CARDOSO, J. A. Direitos Autorais no Trabalho Acadêmico. *Revista Jurídica da Presidência*, v. 9, n. 86, p. 58-86, 2007.

CLARKE, C. L.; CORMACK, G. V.; LASZLO, M.; LYNAM, T. R.; TERRA, E. L. The impact of corpus size on question answering performance. *In*: JÄRVELIN, K.; BEAULIEU, M.; BAEZA-YATES, R.; MYAENG, S. H. (Eds.). *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. New York: Association for Computing Machinery, 2002. p. 369-370.

COSTA, A. S. *Um sistema de anotação de múltiplas camadas para corpora históricos da língua portuguesa baseados em manuscritos*. Doctoral Dissertation (PhD in Linguistics) – Department of Linguistic and Literary Studies (DELL), State University of Southwestern Bahia (UESB), Vitória da Conquista, Bahia, Brazil, 2024.

CRESPO, M. C. R. M.; ROCHA, M. L. S. J.; STURZENEKER, M. L.; SERRAS, F. R.; MELLO, G. L.; COSTA, A. S.; PALMA, M. F.; MESQUITA, R. M.; GUETS, R. P.; SILVA, M. M.; FINGER, M.; PAIXÃO DE SOUSA, M. C.; NAMIUTI, C.; MARTINS DO MONTE, V. Carolina: a General Corpus of Contemporary Brazilian Portuguese with Provenance, Typology and Versioning Information. Manuscript. September 2022. Preprint: arXiv:2303.16098v1 [cs.CL], 28 Mar. 2023.

DAVIES, M.; FERREIRA, M. *Corpus do Português*: 1,1 billion words, Web/Dialetics. Provo, UT: Brigham Young University, 2016. Disponível em: https://www.corpusdoportugues.org/web-dial/. Acesso em: 26 maio 2021.

DAVIES, M.; FERREIRA, M. *Corpus do Português:* 1,1 billion words, NOW. Provo, UT: Brigham Young University, 2018. Disponível em: https://www.corpusdoportugues.org/now/. Acesso em: 26 maio 2021.

DE BRITO, M. G.; VALÉRIO, R. G.; DE ALMEIDA, G. P.; DE OLIVEIRA, L. P. *CORPOBRAS PUC-RIO*: Desenvolvimento e análise de um corpus representativo do português. PUC-Rio, 2007. Disponível em: http://www.puc-rio.br/pibic/relatorio_resumo2007/resumos/LET/marcia_gonzaga_de_brito_rubiae_guilherme_valerio_e_gabriel_paladino_de_almeida.pdf. Acesso em: 26 maio 2021.

DE OLIVEIRA, L. P.; DIAS, M. C. P. Compilação de corpus: representatividade e o CORPOBRAS. *Calidoscópio*, v. 7, n. 3, p. 192-198, 2009. https://doi.org/10.4013/cld.2009.73.03.

DEVLIN, J.; CHANG, M.; LEE, K.; TOUTANOVA, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: BURSTEIN, J.; DORAN, C.; SOLORIO, T. (Eds.). *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*. Cambridge, Massachusetts: Association for Computational Linguistics, 2019. p. 4171–4186. http://dx.doi.org/10.18653/v1/N19-1423.

FERRARESI, A.; BERNARDINI, S.; PICCI, G.; BARONI, M. Web corpora for bilingual lexicography: A pilot study of English/French collocation extraction and translation. *In*: INSTITUTE FOR TRANSLATION STUDIES AND SPECIALISED COMMUNICATION – UNIVERSITY OF HILDESHEIM. *Using Corpora in Contrastive and Translation Studies*. Newcastle: Cambridge Scholars Publishing, 2010. p. 337-362.

FLETCHER, W. H. Concordancing the Web: promise and problems, tools and techniques. In: HUNDT, M.; NESSELHAUF, N.; BIEWER, C. (Eds.). *Corpus Linguistics and the Web*. Amsterdam/New York: Rodopi, 2007. p. 25-45.

JAKUBÍČEK, M.; KILGARRIFF, A.; VOJTĚCH, K.; PAVEL, R.; SUCHOMEL, V. The TenTen corpus family. *In*: 7TH INTERNATIONAL CORPUS LINGUISTICS CONFERENCE CL, 2013. p. 125-127.

KALCHBRENNER, N.; BLUNSOM, P. Two recurrent continuous translation models. *In*: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Cambridge, Massachusetts: Association for Computational Linguistics, 2013. p. 1700–1709.

KILGARRIFF, A.; GREFENSTETTE, G. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*, v. 29, n. 3, p. 333-347, 2003. https://doi.org/10.1162/089120103322711569.

LAN, Z.; CHEN, M.; GOODMAN, S.; GIMPEL, K.; SHARMA, P.; SORICUT, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

LIU, V.; CURRAN, J. R. Web text corpus for natural language processing. *In*: MCCARTHY, D.; WINTNER, S. (Eds.). *11th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2006.

LIU, Y.; OTT, M.; GOYAL, N.; DU, J.; JOSHI, M.; CHEN, D.; LEVY, O.; LEWIS, M.; ZETTLEMOYER, L.; STOYANOV, V. *Roberta*: A robustly optimized bert pretraining approach. 2019. *arXiv preprint arXiv:1907.11692*

NAMIUTI, C.; SANTOS, J. V. Novos desafios para antigas fontes: a experiência DOViC na nova linguística histórica. In: PIMENTA, R. M.; ALVES, D. (Orgs.). *Humanidades digitais e o mundo lusófono*. Rio de Janeiro: Editora FGV, 2021. p. 69-89.

PAIXÃO DE SOUSA, M. C. O Corpus Tycho Brahe: contribuições para as humanidades digitais no Brasil. *Filologia e Linguística Portuguesa*, v. 16, p. 53-93, 2014. https://doi.org/10.11606/issn.2176-9419.v16ispep53-93.

PINHEIRO, G. M.; ALUÍSIO, S. M. *Córpus Nilc*: descrição e análise crítica com vistas ao projeto Lacio-Web. São Paulo: Núcleo Interinstitucional de Lingüística Computacional, 2003. Disponível em: http://143.107.183.175:22180/lacioweb/publicacoes.htm. Acesso em: 27 maio 2021.

RADFORD, A.; NARASIMHAN, K.; SALIMANS, T.; SUTSKEVER, I. Improving language understanding by generative pre-training. *OpenAI*, 2018. Disponível em: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf. Acesso em: 10 jun. 2021.

RAFFEL, C.; SHAZEER, N.; ROBERTS, A.; LEE, K.; NARANG, S.; MATENA, M.; ZHOU, Y.; LI, W.; LIU, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. 2019. *arXiv preprint arXiv:1910.10683*. Disponível em: https://www.jmlr.org/papers/v21/. Acesso em: 27 May 2021.

SALES, Joana; SALES, Teresa. Carolina Michaëlis de Vasconcelos (1851-1925). *Centro de Documentação e Arquivo Feminista Elina Guimarães*, 2025. Disponível em: https://www.cdocfeminista.org/carolina-michaelis-de-vasconcelos-1851-1925/. Acesso em: 25 jan. 2025.

SANH, V.; DEBUT, L.; CHAUMOND, J.; WOLF, T. *DistilBERT, a distilled version of BERT*: smaller, faster, cheaper and lighter. 2019. *arXiv preprint arXiv:1910.01108*

SANTOS, V. G. dos; ALVES, C.; CARLOTTO, B. B.; DIAS, B. A. P.; GRIS, L. R. S.; IZAIAS, R.; MORAIS, M. L.; OLIVEIRA, P.; SICOLI, R.; SVARTMAN, F. R. F.; LEITE, M. Q.; ALUÍSIO, S. M. CORAA NURC-SP Minimal Corpus: a manually annotated corpus of Brazilian Portuguese spontaneous speech. *In*: SPECIAL INTEREST GROUP ON IBERIAN LANGUAGES. *Proc. IberSPEECH 2022*, 2022. p. 161-165. 10.21437/IberSPEECH.2022-33.

SANTOS, D. O projecto Processamento Computacional do Português: Balanço e perspectivas. *In*: DAS GRAÇAS, M. (Ed.). *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)*. São Paulo: ICMC/USP, 2000. p. 105-113.

SANTOS, J. V.; NAMIUTI, C. O futuro das humanidades digitais é o passado. *In*: CARRILHO, E.; MARTINS, A.M.; PEREIRA, S.; SILVESTRE, J.P. *Estudos Linguísticos e Filológicos oferecidos a Ivo Castro*. Lisboa: Centro de Linguística da Universidade de Lisboa, 2019. p. 1381-1404.

SARDINHA, T. B. Lingüística de Corpus: Histórico e Problemática. *DELTA: Documentação de Estudos em Linguística Teórica e Aplicada*, v. 16, n. 2, p. 323-367, 2000. https://doi.org/10.1590/S0102-44502000000200005.

SARDINHA, T. B.; FILHO, J. L. M.; ALAMBERT, E. *Manual Córpus Brasileiro*. São Paulo: PUCSP, FAPESP, 2010. Disponível em: https://www.linguateca.pt/Repositorio/manual_cb.pdf. Acesso em: 26 maio 2021.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. *In*: CERRI, R.; PRATI, R.C. (Eds). *Brazilian Conference on Intelligent Systems*. Cham: Springer, 2020. p. 403-417. Disponível em: https://link.springer.com/chapter/10.1007%2F978-3-030-61377-8_28.

STIM, R. *Fair Use*. Stanford, CA: Stanford Libraries; NOLO, 2016a. Disponível em: https://fairuse.stanford.edu/overview/fair-use/. Acesso em: 27 maio 2021.

STIM, R. *Measuring Fair Use*: The Four Factors. Stanford, CA: Stanford Libraries; NOLO, 2016b. Disponível em: https://fairuse.stanford.edu/overview/fair-use/. Acesso em: 27 maio 2021.

STURZENEKER, M.; CRESPO, M. C.; ROCHA, M. L.; FINGER, M.; PAIXÃO DE SOUSA, M. C.; MARTINS DO MONTE, V.; NAMIUTI, C. Carolina's Methodology: building a large corpus with provenance and typology information. *In*: TROJAHN, C.;

FINATTO, M. J.; PAIVA, V. de; VIEIRA, R. *Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP 2022)*. Aachen, DE: CEUR-WS, Vol. 3128, 2022. Disponível em: http://ceur-ws.org/Vol-3128.

TEI CONSORTIUM, BURNARD, L.; SPERBERG-MCQUEEN, C. M. *TEI P5*: Guidelines for electronic text encoding and interchange. Version 4.8.0. Last updated on 8th July 2024, revision f9891a87. Disponível em: https://tei-c.org/Vault/P5/4.8.0/doc/tei-p5-doc/en/Guidelines.pdf. Acesso em: 4 set. 2024.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. Attention is all you need. *Advances in Neural Information Processing Systems*. 2017. arXiv:1706.03762.

VIANNA, A. E. P. B.; DE OLIVEIRA, L. P. *CORPOBRAS PUC-Rio*: análise de corpus e a metáfora gramatical. Rio de Janeiro: PUC-Rio, 2010. Disponível em: http://www.puc-rio.br/ensinopesq/ccpg/Pibic/relatorio_resumo2010/relatorios/ctch/let/LET-%20Ana%20Elisa%20Piani%20Besserman%20Vianna.pdf. Acesso em: 26 maio 2021.

WAGNER FILHO, J. A.; WILKENS, R.; IDIART, M.; VILLAVICENCIO, A. The BrWac corpus: A new open resource for Brazilian Portuguese. *In*: CALZOLARI, N.; CHOUKRI, K.; CIERI, C.; DECLERCK, T.; GOGGI, S.; HASIDA, K.; ISAHARA, H.; MAEGAARD, B.; MARIANI, J.; MAZO, H.; MORENO, A.; ODIJK, J.; PIPERIDIS, S.; TOKUNAGA, T. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018. p. 4339-4344. Disponível em: https://www.aclweb.org/anthology/L18-1686