Carolina's Methodology: building a large corpus with provenance and typology information

Mariana Lourenço Sturzeneker[0000-0002-0878-3463][1], Maria Clara Ramos Morales Crespo[0000-0003-1500-2749], Maria Lina de Souza Jeannine Rocha[0000-0001-9526-4786], Marcelo Finger[0000-0002-1391-1175], Maria Clara Paixão de Sousa[0000-0002-8422-417X], Vanessa Martins do Monte[0000-0002-4929-5298] and Cristiane Namiuti[0000-0002-1451-8391]

Center for Artificial Intelligence, University of São Paulo. Av. Prof. Lúcio Martins Rodrigues, 370 - 05508-020 - Butantã, São Paulo, Brazil.
lavihd@usp.br

Abstract. This paper presents the salient aspects of WaC-wiPT methodology, developed for the construction of the Carolina Open Corpus for Linguistics and Artificial Intelligence, a large corpus for contemporary Brazilian Portuguese. Both the corpus and the methodology are under development at the Center for Artificial Intelligence of the University of São Paulo. This paper describes the paths we took this far into the making of the Carolina Corpus, presents its current state and discloses the future agenda of the project.

Keywords: Open Corpus, Brazilian Portuguese, Provenance, Typology.

1 Introduction

The Carolina Open Corpus for Linguistics and Artificial Intelligence is a general corpus of contemporary Brazilian Portuguese texts written after 1970 hosting provenance and typology information. It is under development since September 2020 as part of the Natural Language Processing for Portuguese (NLP2) project of the Center for Artificial Intelligence of the University of São Paulo (C4AI-USP).

With Carolina, we expect to build a large and reliable resource for research in both Linguistics and Computer Sciences, with more than a billion tokens. In opposition to other corpora built under the "Web as corpus" view, which aim to gather large amounts of texts for language-modeling retrieving them from multiple untraceable origins, Carolina's intention is to curate sources in large quantities of text. In doing that, we expect to provide further information about the texts, especially on provenance and typology [1], which benefit linguistic research.

By not tailoring the corpus for any specific linguistic application, we aim to avoid restraining the possibilities for future projects and posing obstacles for any researchers interested in studying a wide range of language aspects. For instance, investigation of typological characteristics, word collocation, language detection and Historical Linguistics. To reach this goal, we developed the WaC-wiPT (Web-as-Corpus with Provenance and Typology) methodology, which combines the automation and large extension of language-modeling corpora with the careful text-information curatorship of smaller linguistic corpora.

2 Related works

Over the last decades, corpus-building initiatives have increasingly resorted to the Web as their main source. As part of this endeavor, the WaCky (Web-As-Corpus Kool Yinitiative) methodology was developed [2, 3, 4]. As it proved to be a relatively easy and not high-resource-demanding method, this framework quickly became popularized. There have been undertakings to apply it to the Portuguese language, such as the Brazilian Portuguese Web as Corpus (brWaC), considered to be the “biggest Brazilian Portuguese corpus available” at the time [5], with 2.68 billion tokens.

One example of a Brazilian Portuguese corpus with a significant size is the Brazilian Corpus [6], with approximately one billion syntactically annotated words. There are also other important corpora out of the envisioned scope of language or size, such as the Oscar Corpus [7], or the Corpus do Português: Web/Dialects [8]. Other corpora with provenance and typology information are known, such as ReLi [9] and CETENFolha [10], albeit their smaller sizes according to their specific goals.

3 Methodology

During the first stages of our research, some effort was put into investigating the possibility of implementing pre-existing corpora-building frameworks, such as the WaCky methodology [2]. However, it does not ensure the transparency of the content that is scraped, so a post-hoc investigation is necessary [2]. This poses a challenge for provenance tracking, quality control, and rights-of-use compliance, which are at the core of Carolina's objectives. Therefore, we built on the knowledge provided by this investigation to develop WaC-wiPT, a Web-as-Corpus with Provenance and Typology methodology, which is constantly being improved.

3.1 Web text prospection

The fundamental steps of the method are based on broad types of domains, the *Carolina broad typology*, which are not intended to reflect the textual content of our documents, but a macro-structure that guides the corpus development. They were defined after the surveying process described below. Currently, we are working with eight: *datasets and other corpora*, *Brazilian Judicial branch* and *Legislative branch*, *journalistic texts*, *public domain works*, *social media*[2], *university domains* and *wikis*. The Carolina broad

[2] It is important to clarify that we only incorporated open access content, in compliance with each domain’s license.

typology opposes the *Carolina narrow typology*, which aims to effectively reflect the textual types of our documents, which are not yet defined for they require further analysis of the texts and may depend on linguistic theory; as well as the *source typology*, a simple typological organization that reflects exclusively what is declared on the sources and therefore are declared when available.

We began by conducting prospective surveys, which are in-depth research of each Web domain, prioritizing open access content available online. In these surveys, we verified if the texts were in our scope, in addition to searching for metadata information and mapping the basic directory structure of each domain. Therefore, this first step is important to help us systematize metadata for future automatic annotation — so we do not lose any important information from our sources — and organize the provenance of the data as well. These surveys also facilitate the download and extraction stages, for they allow the downloaded content to be mostly deliberate and not randomly crawled.

At the beginning of the data collection step, we mirrored some websites to keep raw copies of our sources, but even in those cases, we could directly extract the desired texts because of the mapping of the directory structure previously made. However, most websites were not obtained with this method for some had defense mechanisms against machine download and others could only be used partially, as they contained texts that were not in our scope or were under restrictive licenses. In all cases, special care was taken to verify the rights of use: during the data collection step, we downloaded exclusively open access texts and later verified if they allowed derivative works. Should the occasion arise that any data is copyright-claimed, our methodology enables the easy removal of any set of texts from the corpus.

3.2 Metadata and extraction

As we aim to build a corpus with provenance and typology information, each text is embedded in an XML header carrying annotated metadata — such as source URL and license — following the TEI (Text Encoding Initiative) guidelines. To gather information for the header, after the download stage some surveys are complemented by opening a small sample of the raw documents and searching for any additional metadata. However, as we have a large number of texts, information that cannot be automatically annotated is not mandatory, thus, most categories are optional, such as Author and Regional Origin. To make this process easier we extract the texts by batches arranged by coincident information, usually grouped by downloaded directory structure, which mirrors the broad typology's.

This means that for the extraction of a batch, we inform the metadata that holds for the whole set, prioritizing the fulfilling of the header’s mandatory categories. Therefore, the metadata collection and insertion are carefully made, to prevent errors from being repeated in the whole batch. We centralize this latter process with an extraction module developed in Python3, which obtains some metadata by input and others automatically, organizes them and generates the XML file with the clean text embedded in the header[3]. It also verifies if the text is valid by assessing its language and size, for example. In addition to the traditional search by words, the structured header allows searches by tags, facilitating the metadata recovery, thus providing further query tools.

In order to test our tools, we processed a portion of our raw data before our first official extraction, with over a billion tokens total and about 24 hours of CPU time. This first test version will not be made publicly available; however, it sheds some light on what is to be expected of the first official publication in terms of size and typology distribution. The texts obtained were sampled randomly (590 files in total) by Carolina broad typology, of which only 4 broad types were included.

We carefully looked for problems concerning not only the cleaning process, but also the metadata provided, and implemented computational solutions to improve textual quality, such as removing remaining blank lines and corrupted characters. Some recurring issues were over or under-cleaned texts, as well as the formatting of the data provided automatically, which sometimes did not match our chosen standards. After this process, we established the importance of human inspection of the files, as some problems would not be easily identified and fixed without it. Therefore, this method of examining the samples of the extracted files will be kept in the future as well as some machine inspections will be made to verify if all the files are well-formatted.

4 Current state

At present, Carolina is at a prototypical stage. After the test extraction, some considerable yet expected size-reduction took place in relation to the raw crawled content[4]. The main reason for that is that files were disposed of tags and non-content elements aiming at a text as clean as possible, and many files did not reach a significant number of characters and were thus discarded. The table ahead illustrates these reductions.

Table 1. Test extraction results.

Broad typology[5]	**Size (GB)**		**Number of words**		**CPU time (h)**
	Raw	**XML files**	**Raw**	**Clean text**	
Judicial branch	71	1,7	2.543.098.232	191.635.110	20,32
Datasets and other corpora	31	7,5	2.870.318.559	327.927.677	0,72
Public domain works	0,17	0,024	4.774.114	3.170.682	0,23
Wikis	741	20	89.664.268.580	665.638.761	3,00
Total	843,17	29,224	95.082.459.485	1.188.372.230	24,27

[3] An example of a generated XML header can be accessed at: https://sites.usp.br/corpuscarolina/exemplo/

[4] The raw crawled content encompasses everything downloaded from each domain, including media files, source metadata and content out of the project's current scope.

[5] A list of the content available on the first version of the Carolina Corpus can be accessed at: https://sites.usp.br/corpuscarolina/repositorios/

Based on this initial testing of our methods' performance, we have already begun the proceedings for the first version, to be released in March 2022, the Carolina 1.0 (Ada), following the steps of this under development methodology.

5 Conclusion and future steps

Both the preliminary surveys and the close analysis of the samples extracted for the test version proved to be essential to keep track of the metadata collection and sustain such complex text headers. Thus, our methodology helped to guarantee the provenance and typology information we aim to preserve in the automatic processes of text extraction.

As for future intentions, we wish to develop language verification tools capable of distinguishing Brazilian Portuguese from other variants and assess the percentage of all languages used in a text, hoping that would allow studies on language contact and cultural influences. Additionally, there are plans to build a historical corpus for Philological and Historical Linguistics research based on our methodology.

For the future versions of the Carolina corpus, we will work to ensure better balancing of text types, which require further effort put into new surveys. In our understanding, this continuous development of the methodology is an essential part of the labor involved in the construction of such an ever-growing corpus.

References

1. Finger, M., Paixão de Souza, M. C., Namiuti, C., Monte, V. M., Costa, A. S., Serras, F. R., Sturzeneker, M. L., Guets, R. P., Mesquita, R. M., Crespo, M. C. R. M., Rocha, M. L. S. J., Palma, M. F., Silva, M. M., Brasil, P. Carolina: a General Corpus of Contemporary Brazilian Portuguese with Provenance and Typology Information. Language resources and evaluation, submitted paper (2021).
2. Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E.: The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. Language resources and evaluation, 43(3), 209-226 (2009).
3. Bernardini, S., Baroni M., Evert, E.: A WaCky introduction. In: Baroni, M., Bernardini, S. (eds.) WaCky! working papers on the web as corpus, pp. 9-40. GEDIT, Bologna (2006).
4. Ferraresi, A., Bernardini, S., Picci, G., Baroni, M.: Web corpora for bilingual lexicography: A pilot study of English/French collocation extraction and translation. In: Using Corpora in Contrastive and Translation Studies, pp. 337-362. Cambridge Scholars Publishing, Newcastle (2010).
5. 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.) PROPOR 2014. LNCS, vol. 8775, pp. 201–206. Springer, Heidelberg (2014).
6. Sardinha, T. B., Filho, J. L. M., Alambert, E.: Manual Córpus Brasileiro, https://www.linguateca.pt/Repositorio/manual_cb.pdf, last accessed 2021/12/13.
7. Suárez, P. J. O. S., Sagot, B., Romary, L.: Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In: Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019, pp. 9-16. Leibniz-Institut für Deutsche Sprach, Mannheim (2019).

8. Davies, M., Ferreira, M.: Corpus do Português: Web/Dialetics, https://www.corpusdoportugues.org/web-dial/, last accessed 2021/12/13.
9. Corpus ReLi, https://www.linguateca.pt/Repositorio/ReLi/, last accessed 2022/03/10.
10. CETENFolha, https://www.linguateca.pt/CETENFolha/, last accessed 2022/03/10.