The construction of a corpus from the Brazilian Historical-Biographical Dictionary

Lucas Ribeiro[1 [0000−0002−8693−2006]], Jaqueline P. Zulini[2], and Alexandre Rademaker[1,3 [0000−0002−7583−0792]]

[1] FGV/EMAp, Brazil
[2] FGV/CPDOC, Brazil
[3] IBM Research and FGV/EMAp

Abstract. We present our ongoing efforts towards the creation of a new Portuguese corpus based on the "Dicionário Histórico-Bibliográfico Brasileiro". The aim to add as many linguistic annotations as possible using widely accepted annotation schemas and distributing all data in standard formats. This first exploratory work revisits what is already done and tests different tools to detect errors and look for the best methods to tackle the problem. Data is available at https://github.com/cpdoc/dhbb-nlp, and it will be continuously improved.

Keywords: Portuguese corpus · segmentation · natural language processing

1 Introduction

In this paper we present our ongoing efforts towards the creation of a corpus based on the "Dicionário Histórico-Bibliográfico Brasileiro" (DHBB) [1] that contains almost 12 millions tokens in about three hundred thousand sentences. The DHBB is a reference work, written by historians and social scientists, it contains almost eight thousand entries with information ranging from the life and career trajectories of individuals to the relationships between the characters and events that the country has hosted. This work presents an exploratory approach in order to find the best methods to create a reliable corpus.

In previous articles, some initial efforts to mine the DHBB texts were presented. In [15], the documents were processed by Freeling [13] to annotate named entities and expand the coverage of OpenWordnet-PT [14]. In [9], DHBB was syntactical analysed with two different statistical parsers. Once a preliminary set of relevant entities types and semantic relations was identified, the classification of appositives syntactic annotations regarding the semantic relations was performed in a sample and evaluated. The low agreement on the appositives annotations between the two parsers shown that results are not very reliable. Later on, in [10], DHBB was subject to syntactical analysis by PALAVRAS

parser [3] and semantic annotated by AC/DC [5].[4] The problem is that not all PALAVRAS errors were fixed but only proper names segmentation. Besides that, the annotation schema used is not widely adopted.

Our aim is to annotate DHBB (Section 2) with as many as possible layers of linguistic information using widely accepted schemas (Section 3) shared by the NLP community. This work is also leading us to improve the DHBB itself, since we could detect several problems like encoding issues and typos in the corpus. We start dealing with the problem of text segmentation into sentences (Section 4). Due to its format, there are several uncommon abbreviations and quotes, thus turning the segmentation task not trivial. By comparing the segmentation of two different tools, we found the divergent results that suggest possible hard cases. Going further, in Section 5, we made a first preliminary experiment on part-of-speech (POS) tagging using the same tool trained on two different corpora. Again, by comparing the outputs, we could make a first confusion matrix and locate where the tools diverge. We also note that comparing the outputs of two different models searching for weakness or inconsistencies in annotation is known strategy, used for instance in [19]. We extend this strategy to segmentation task as well.

2 DHBB

DHBB [1] is an encyclopedia developed and curated by Centro de Pesquisa e Documentação de História Contemporânea do Brasil (CPDOC), from Fundação Getulio Vargas (FGV), and is an important resource for all research, nationally and internationally, interested in Brazilian politics. It was first published in 1984, in four volumes containing 4,500 entries. In the 2001, the resource was increased by one more volume reaching a total of 6,620 entries, and in 2010 its material was made available online,[5] with about 7,500 entries composed of a title, the kind of entry (biographical or thematic), the author of the entry, and the text in a text field. Currently, DHBB has 7,687 entries and data is maintained in text files under version control.[6] The process and rationale of releasing this content from the database and converting it to full text aiming at natural language processing are described by [17]. Each entry became a single text file that received a unique identifier, and new metadata were added, such as the gender of the biographee and the political role she/he had.[7]

As noted before, DHBB was already treated by AC/DC, using PALAVRAS [8] to parse it. However the only mapping for PALAVRAS to other schemas is the

[4] Data available at https://www.linguateca.pt/acesso/corpus.php?corpus=DHBB. The AC/DC is a online service for corpus browsing and searching but also a data format with attributes added on tokens in the output of the PALAVRAS parser.

[5] https://cpdoc.fgv.br/acervo/dhbb

[6] https://github.com/cpdoc/dhbb

[7] In order to preserve the original repository, the present work can be reproduced from https://github.com/cpdoc/dhbb-nlp where all data is available.

[8] The documentation is available at https://visl.sdu.dk/visl/pt/info/.

one to Universal Dependencies used by [16] but outdated. A corpus with syntactic annotation on a non-widespread schema is less useful. In a language processing pipeline, we usually aspire, for example, that a named entity classifier component can understand the syntactic annotation of the parser, and both results could be easily combined for a further processing step. Particular syntactic annotations force the implementation of mappings not always full content preserving. Furthermore, AC/DC has joint DHBB alongside with other two corpora, the 'Dicionário histórico-biográfico da Primeira República' (DHBPR) and the 'Dicionário da política republicana do Rio de Janeiro'. Our aim is just to deal with the DHBB which is already a big corpus.

Finally, we mention that the texts in DHBB were written with certain independence among their writers, and this directly impacts any computational approach in the corpus. For instance, abbreviations and punctuations were not used consistently. Moreover, some authors may decide to cite another text in the DHBB, whereas others do not, thus leading to a more clean text, without citations or references.

3 Universal Dependencies for Portuguese

As stated in [11], the Universal Dependencies project provides an inventory of dependency relations that are linguistically motivated, computationally useful, and cross-linguistically applicable. It holds two Portuguese corpora since its 2.1 release. The UD Portuguese Bosque corpus (Bosque) [16] is a subset of the Floresta Sinta(c)tica treebank [7] converted to UD annotation style. Bosque currently has 9,365 sentences that were taken from CETENFolha (4,213 sentences in Brazilian PT) and CETEMPublico (5,152 sentences in European PT) corpora. The other Portuguese corpus is the Google Stanford Dependencies (GSD) [12]. This corpus was converted from the Google Universal Dependency Treebank, but we were unable to find information about the origin of the data. Inspecting the sentences manually, we found pieces of evidence of being text collected from newspaper articles, probably between 2010 and 2012.

It is important to emphasize that DHBB and the UD Portuguese corpora have different text styles. DHBB text has an encyclopedia-style; many sentences do not have a subject since, in the same entry, we are usually talking about the same entity. DHBB authors followed a general guideline for making sentences declarative and neutral. The guidelines also suggest the flow of information on each entry (i.e., when describing a politician, the first paragraph contains the city and year of birth and the parents' names, while the second paragraph details the education path, and so on). On the other hand, news from different sources tends to bias convincement and impact on the readers. Vocabulary is much more assorted compared to DHBB, given they are not limited to the history of politicians but cover many domains from sports to science.

4 Text Segmentation

Since manually revising more than three hundred thousand sentences is almost impractical, we adopted a strategy already suggested in the literature. Using two different tools, we segmented DHBB and compared the results, revising the cases where the tools diverge. In that way, we focused on systematic confusions, an approximation for the hard cases that should be manually inspected by humans. The Apache OpenNLP [6] is a machine learning based toolkit for natural language processing. The sentence detector module uses a maximum entropy model to evaluate end-of-sentence characters to determine if they signify the end of a sentence. Sentence Detection can be done before or after tokenization.[9] One important limitation of OpenNLP sentence segmentation module is that it cannot identify sentence boundaries based on the contents of the sentence. A notable example is the first sentence in the articles where the title is mistakenly identified to be the first part of the first sentence.

Freeling [13] is an open-source multilingual processing library providing a wide range of analysis functionalities. Its sentence splitter module receives a list of word objects and returns a list of sentence objects. It uses an options file, where one can tune parameters in order to fit to particularities of the text in question.

In Linguateca processing steps of DHBB,[10] before DHBB was syntactically analyzed with PALAVRAS, the corpus was tokenized and segmented with a Perl library called `Lingua::PT::PLNbase`. [11] As stated on its website, this library was created in 2004, and the main difference of this library compared to Freeling is that it has fewer configuration options, so users have less control over the tokenization and segmentation steps. For instance, Linguateca's tokenizer cannot deal appropriately with characters such as 'º', like *"(...) em seu artigo 14, parágrafos 10º ..."* and has to replace these characters with ASCII symbols, thus making the outputs different from the inputs. We didn't find how to add new abbreviations to the list handled by the library, as we did for Freeling. Table 1 shows the number of sentences of the DHBB parsed with Linguateca, OpenNLP and Freeling.

Table 1. Number of sentences of the DHBB with each parser.

Tool	Number of Sentences
Linguateca	312,539
OpenNLP	314,930
Freeling	311,530

We first processed all DHBB files with Freeling and OpenNLP. At first, we found that 3,438 files diverged on segmentation, comprising 44% of the whole

[9] Here we have used the pre-trained models that adopted segmentation first.
[10] Documented at https://www.linguateca.pt/acesso/anotacao.html.
[11] See https://metacpan.org/pod/Lingua::PT::PLNbase.

DHBB. Looking at the divergent files,[12] we could divide the errors in two groups: (i) problems in the DHBB itself; (ii) confusion of the tools due to uncommon abbreviations, quotations that lasts longer than a sentence, and uncommon quotation symbols.

Comparing the output of the tools was very efficient to find errors on the corpus. We found many occurrences of the unicode non-breaking space character (wrongly in place of normal spaces) in more than 1,700 files. This character confuse both tools in different contexts. During the migration of DHBB data to text files, some sentences have been split into more than one line and some segments were lost, ending up with sentences fragments without a final period. Freeling is able to detect sentences regardless of the presence of line breaks, but OpenNLP is not, it never join separated lines into one sentence. We have fixed many cases and listed some for further review by DHBB editors. Another error is related to many different quotation marks, some not properly balanced. In some cases, we found even hard for humans to identify the quotes given the nesting of quotation marks. Finally, during the migration of DHBB to text files, the titles of the entries are moved from the first line of the text to a metadata field, but many cases remain in the text. We manually cleaned up around 165 files.

Regarding the divergence of the tools, this happens due to uncommon abbreviations, quotations and punctuation symbols. For quotes, Freeling has a fine-tuned control for blocking or not the introduction of sentence split inside a pair of parenthesis-like markers. These markers can be quotes, parenthesis or any other pair of characters. OpenNLP expertise is limited by the data it saw during training. OpenNLP sentence detector was trained with the CoNLL-X shared task [4], the Portuguese part is taken from the Bosque corpus distributed by Linguateca before its conversion to UD (Section 3).

Due to DHBB political content, abbreviations of names and initials of political parties are very common on it. We have many occurrences of PP (Partido Progressista), PT (Partido dos Trabalhadores) and names such as 'M. H. Simonsen'. In Freeling, we have some ways to control the segmentation. First, we were able to refine the list of abbreviations that must not be separated of their following dot during tokenization. Since tokenization happens before sentence detection, dots part of abbreviations are not considered candidates for sentence ending characters. When the abbreviation happens in the end of the sentence, a sentence split will only be introduced by Freeling if a sentence ending character is followed by a capitalized word or a sentence start character. That is, a sentence ending with an abbreviation followed by a sentence that starts with a proper name will always confuse Freeling. OpenNLP tends to correctly detect sentences that starts with a quotation mark or a non-standard or multi-character sentences ending marks such as colon, semicolon and ellipsis. Finally, we observed that although we could have defined quotation marks as possible sentence start character for Freeling, this option increase the number of errors in our experiments. Regarding OpenNLP training, we found that abbreviation dictionary can

[12] We have used the `diff` command line tool inside Emacs editor. This environment provided us a easly way to inspect the differences in the files in an interactive manner.

also be provided but many details regarding the parameters used by the module are not well-documented.

After we calibrated Freeling parameters, we improved the results and obtained that of the 7,687 files, 2,096 diverged on the number of sentences, comprising 27% of the total. We note that, despite the huge amount of divergent files, when counting the number of divergent sentences, they comprise a total of 2% of the total. Of course, we are aware that we may still have cases of false negative, that is, when the tools agree in a wrong segmentation.

Once we identified the differences between Freeling and OpenNLP, we processed the corpus with `Lingua::PT::PLNbase`. A detailed exploration of the differences was not yet possible due to the following problems. First, the encoding of some files was lost with the mixing of UTF-8 and ISO-8859 code systems. Next, XML tags were included to mark the begin/end of sentences, but the tool did not produce valid XML files (i.e. symbols such as '&' were not converted to '&' and the markup was not a correct tree structure). These errors make difficult the processing of the outputs by an XML parser. Given the results that we had from `Lingua::PT::PLNbase`, we could only count the number of sentences but we could not easy compare the differences of the files.

Finally, we emphasize that this is an initial exploratory approach to decide what methods are the best to tackle the corpus analysis, and also this initial efforts could detect problems in the DHBB itself, providing an improvement of the whole corpus, which will reduce future errors and contribute to make DHBB more reliable.

5 Part-of-speech tagging

After tokenization and segmentation, next step will comprise the part-of-speech (POS) tagging. Following the same approach, here we briefly explain our first experiment sharing preliminary results. In order to identify possible errors in POS tagging, we compared the outputs of two different UDPipe [18] models trained on Bosque and GSD corpora when applied to DHBB. As a first approach we made a confusion matrix, shown in Table 2. The rows are GSD tags and the columns are from Bosque. Similar technique was employed in [8].

As a first glance, we found many repeated errors. For example, GSD model tag months as NOUN whereas Bosque model tags them as PROPN. Since many recent revisions of Bosque driven by changes in the UD guidelines [16] were not yet applied to GSD corpus, many differences on the results of the models are expected. Moreover, the difference between the vocabulary and the style between the training data (UD corpora) and DHBB, and the strategy of UDPipe on dealing with out-of-vocabulary words explain a lot of differences too. These will all be subject of future work.

It is worth to explain the choice of UDPipe for this first experiment with POS tagging. UDPipe is easier to run/train using corpora annotated with UD tags and following UD guidelines, and it produces UD compatible output. The POS tagger of Freeling does not follow the UD guidelines, and we haven't yet

Table 2. Confusion Matrix between GSD and Bosque

	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PART	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X
ADJ	74193	226	270	14	3	369	0	4845	32	0	90	882	1	0	0	4137	7
ADP	88	409490	683	12	0	1500	0	169	3	0	132	952	0	59	0	190	7
ADV	990	635	41643	96	96	41	0	1609	9	0	113	273	0	15	0	95	0
AUX	109	4	45	36417	3	15	0	112	0	0	18	39	0	0	0	5036	0
CCONJ	9	68	1352	2	50222	12	0	25	0	0	450	107	1	4571	0	26	1
DET	579	3099	92	0	0	293250	0	283	316	0	1134	195	0	0	0	123	0
INTJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NOUN	5632	324	921	81	1	160	1	293786	205	0	234	2294	0	14	0	3447	14
NUM	76	182	67	0	0	260	0	346	72184	0	3	474	10	7	0	9	1
PART	13	5	8	0	0	0	0	20	1	0	300	27	0	9	0	6	0
PRON	71	85	48	3	0	558	0	128	1357	0	22890	276	0	970	0	19	5
PROPN	3785	18470	1396	5441	60	1254	9	56300	133	0	647	300078	89	171	39	2476	1
PUNCT	101	4	0	0	1	0	0	611	0	0	6	145	268560	0	36	0	0
SCONJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SYM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	52	0	0
VERB	2710	211	656	4771	39	50	0	4695	128	1	54	444	1	11	0	154572	3
X	46	1	2	0	0	11	0	14	0	0	6	12	0	0	1	11	0

explored the options for training and evaluating the OpenNLP tagger. Moreover, our focus is not on the adaptation of a single tool for processing DHBB, but on the production of consistent data with compatible annotations.

6 Conclusion

As observed by [2], tokenization is a crucial component of language processing, yet there is no widely accepted tokenization method for English texts. In this article, we have shown that the segmentation of texts into sentences is also not a solved problem in general, as many researchers believe. Many language processing tools are trained or fine-tuned to deal with news articles. Still, once they are used to process texts with a different narrative style (such as encyclopedia material) or domain-specific documents, many issues appear unsolved.

The observation that motivates our work is that the current version of DHBB in AC/DC is not complete since (i) the tool used to segment the corpus is not reliable; (ii) the output format of the parsed files is not widely used. Moreover, analyzing the divergence on segmentation led us to detect errors in the corpus itself, thus leading to an improvement of the DHBB.

As we go deeper in the annotation of DHBB, we note that there are several issues that need to be solved in order to properly annotate such important material. We started with segmentation and noticed that relying just on one tool can lead to wrong segmentation, therefore adding different tools is a good technique to identify and correct possible errors, reducing the manual revision effort. We also note that this investigation can lead us to find weakness or inconsistencies on corpora used to train the models for POS tagging and parsing.

After this exploratory analysis, future work will need to be done. For instance, we plan to retrain the OpenNLP sentence detector with fragments of the DHBB manually curated and also add training UDPipe models with manually annotated DHBB texts in order to look for improvements in the POS tagging.

We also note that although the use of machine-learning tools in natural language processing is prevalent, the generality of the pre-trained models is rarely discussed in the literature. The openNLP segmentation model that we used proved to be not well adapted to DHBB style and maybe overfit to the data used for creating it, but training a model with DHBB data would not also make the result a robust model for processing other materials. In that sense, we may be still far from advances in general and reliable language processing techniques.

References

1. de Abreu, A., Lattman-Weltman, F., de Paula, C.J. (eds.): Dicionário Histórico-Biográfico Brasileiro pos-1930. CPDOC/FGV, Rio de Janeiro, 3 edn. (2010), http://cpdoc.fgv.br/acervo/dhbb
2. Barrett, N., Weber-Jahnke, J.: Building a biomedical tokenizer using the token lattice design pattern and the adapted viterbi algorithm. BMC bioinformatics **12**(3), S1 (2011)
3. Bick, E.: The parsing system palavras. Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework (2000)
4. Buchholz, S., Marsi, E.: CoNLL-x shared task on multilingual dependency parsing. In: Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X). pp. 149–164. Association for Computational Linguistics, New York City (Jun 2006), https://www.aclweb.org/anthology/W06-2920
5. Costa, L., Santos, D., Rocha, P.A.: Estudando o português tal como é usado: o serviço ac/dc. In: quot; In The 7 th Brazilian Symposium in Information and Human Language Technology (STIL 2009)(São Carlos Brasil 8-11 de Setembro de 2009) (2009)
6. Foundation, A.S.: Apache opennlp. version 1.9.1. (2019), https://opennlp.apache.org
7. Freitas, C., Rocha, P., Bick, E.: Floresta sintá (c) tica: bigger, thicker and easier. In: International Conference on Computational Processing of the Portuguese Language. pp. 216–219. Springer (2008)
8. Freitas, C., Trugo, L.F., Chalub, F., Paulino-Passos, G., Rademaker, A.: Tagsets and datasets: Some experiments based on portuguese language. In: International Conference on Computational Processing of the Portuguese Language. pp. 459–469. Springer, Canela, RS, Brazil (2018), https://rd.springer.com/book/10.1007/978-3-319-99722-3
9. Higuchi, S., Freitas, C., Cuconato, B., Rademaker, A.: Text mining for history: first steps on building a large dataset. In: Proceedings of 11th edition of the Language Resources and Evaluation Conference. Miyazaki, Japan (May 2018), http://www.lrec-conf.org/proceedings/lrec2018/summaries/1084.html
10. Higuchi, S., Santos, D., Freitas, C., Rademaker, A.: Distant reading brazilian politics. In: Navarretta, C., Agirrezabal, M., Maegaard, B. (eds.) Proceedings of the Digital Humanities in the Nordic Countries 4th Conference. vol. 2364. Copenhagen, Denmark (Mar 2019), http://ceur-ws.org/Vol-2364/
11. Jurafsky, D., H. Martin, J. (eds.): Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Stanford, 3 edn. (2019), https://web.stanford.edu/ jurafsky/slp3/

12. McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Oscar, T., et al.: Universal dependency annotation for multilingual parsing. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 92–97 (2013)
13. Padró, L., Stanilovsky, E.: Freeling 3.0: Towards wider multilinguality. In: LREC2012 (2012)
14. de Paiva, V., Rademaker, A., de Melo, G.: Openwordnet-pt: An open Brazilian Wordnet for reasoning. In: Proceedings of COLING 2012: Demonstration Papers. pp. 353–360. The COLING 2012 Organizing Committee, Mumbai, India (Dec 2012), http://www.aclweb.org/anthology/C12-3044, published also as Techreport http://hdl.handle.net/10438/10274
15. Paiva, V.D., Oliveira, D., Higuchi, S., Rademaker, A., Melo, G.D.: Exploratory information extraction from a historical dictionary. In: IEEE 10th International Conference on e-Science (e-Science). vol. 2, pp. 11–18. IEEE (Oct 2014). https://doi.org/http://dx.doi.org/10.1109/eScience.2014.50
16. Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., de Paiva Universal Dependencies for Portuguese, V.: Universal dependencies for portuguese. In: Proceedings of the Fourth International Conference on Dependency Linguistics (Depling). pp. 197–206. Pisa, Italy (Sep 2017)
17. Rademaker, A., Oliveira, D.A.B., de Paiva, V., Higuchi, S., e Sá, A.M., Alvim, M.: A linked open data architecture for the historical archives of the getulio vargas foundation. International Journal on Digital Libraries **15**(2-4), 153–167 (2015). https://doi.org/10.1007/s00799-015-0147-1, http://dx.doi.org/10.1007/s00799-015-0147-1
18. Straka, M., Straková, J.: Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. pp. 88–99. Association for Computational Linguistics, Vancouver, Canada (August 2017), http://www.aclweb.org/anthology/K/K17/K17-3009.pdf
19. Volokh, A., Neumann, G.: Automatic detection and correction of errors in dependency tree-banks. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. pp. 346–350. Association for Computational Linguistics (2011)