# Automatic information extraction: a distant reading of the Brazilian Historical-Biographical Dictionary

Suemi Higuchi[1[0000-0002-6255-3781]*], Claudia Freitas[2[0000-0001-6807-8558]] and Diana Santos[3[0000-0002-3108-7706]]

[1] Fundação Getulio Vargas, Rio de Janeiro, Brazil
[2] Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Brazil
[3] Linguateca & University of Oslo, Oslo, Norway
suemi.higuchi@fgv.br, claudiafreitas@puc-rio.br, d.s.m.santos@ilos.uio.no

**Abstract.** We present some results of applying natural language processing (NLP) techniques in the domain of History, having as object of investigation the Brazilian Historical-Biographical Dictionary (*Dicionário Histórico-Biográfico Brasileiro*, DHBB). After improving or adding annotation of specific fields, information extraction techniques based on manually derived patterns were applied to three relevant problems: the age of entrance in Brazilian politics, the academic background of Brazilian politicians, and family ties among the political elites.

**Keywords:** corpus linguistics, information extraction, distant reading, Brazilian politics

## 1 Introduction

*Dicionário Histórico-Biográfico Brasileiro* (Brazilian Historical-Biographical Dictionary or DHBB for short), an encyclopedic work conceived by Centro de Pesquisa e Documentação de História Contemporânea do Brasil (CPDOC) from Fundação Getulio Vargas (FGV), gathers more than 7,500 biographical and thematic entries on the contemporary history of Brazil, and contains information ranging from the life trajectory, education and career of individuals, to the relationships built between the characters and events that the country hosted [1].

Our goal with the present work was to create, from the DHBB, an annotated corpus for automatic information extraction's purpose, enabling "distant readings" of Brazilian contemporary political history[7]. In 2018, the DHBB corpus was thus integrated into the AC/DC[1] collection and can be freely consulted by the linguistic and NLP communities. In its present version 7.3, it contains 457,101 sentences, almost 16 million tokens and about 14 million words. As it is an organic work, in constant updating, new versions of the corpus may be created and released from time

* * Author to whom correspondence should be addressed.

1 Available at https://www.linguateca.pt/acesso/corpus.php?corpus=DHBB.

to time. Furthermore, DHBB is also available to the research community in an open repository under version control[2][8].

The complete process of creating the DHBB corpus includes the morphosyntactic analysis of the material, the identification of domain relevant entities, the addition of semantic annotation to the corpus, the definition of semantic relations of interest and the mapping of lexical-syntactic patterns expressing these relations. These steps prepare the texts for the identification of the structures of interest, which are then extracted and presented in a structured way. We evaluate here a set of textual patterns according to their productivity in the DHBB, for the following topics: age of the politicians when entering public life, their academic training and family ties.

Our assumption is that by using specifically designed lexical-syntactic patterns it is possible to extract high quality information from an annotated corpus, at least in an encyclopedic genre in the History domain.

The main motivation to explore the DHBB through computational linguistic tools arose from the need to search for certain kinds of information without closely reading a large number of entries. A survey carried out with researchers who frequently consult the Dictionary asked them which questions they would have liked to have answered automatically. From this survey, questions emerged such as: How is the educational background of political cadres characterized over time? How old were Supreme Court ministers when they were nominated? Who are the politicians who have family ties to other politicians? Which ties?

The answers to these questions are found dispersed in the entries and are not indexed in metadata fields. An information extraction system (IE) helps to overcome this challenge, as it aims to select and obtain specific information from large volumes of text.

## 2 Information Extraction

When we talk about information extraction (IE) we are referring to the process of automatically obtaining structures – such as entities, relationships between entities and attributes that describe entities – from unstructured sources. An IE system can, for example, identify and extract from a collection of texts the name of all the mentioned organizations - including those that the user had no prior knowledge of -, the name of all the people who have any link with these organizations, and the type of link.

Among the strategies adopted in these systems, there are those that use inference rules from linguistic clues and those that use lexical lists. The clues enable the identification of patterns and can be based on different features, such as morphosyntactic, orthographic, context and so on, and may be language dependent or not. Lexical lists are simpler and do not consider the contexts in which the terms appear; one of the motivations for its use is the fact that parsers are not able to identify with 100% accuracy certain entities, even with the help of morphosyntactic heuristics applied in named entity recognition (NER). As for computational methods, we can roughly classify them into rule-based approaches and machine learning

[2] Available at https://github.com/cpdoc/dhbb

approaches (supervised, semi-supervised and unsupervised), and there are those that combine both.

Since we use the first approach, we are only going to describe it here, not the second. Extraction rules are developed manually and mainly relies on lexical-syntactic aspects of sentences. Although at first sight they are constructions that are presented in a very simple way in the language, the formalization of these patterns brings positive results as they serve as clues for the automatic discovery of information structures in the text. According to Hearst [5], different relationships can be expressed using a small number of lexical-syntactic patterns. Her work to extract *is-a* relationships is widely known and cited. These patterns are constructed from phrases containing clues such as "and", "such as", "like", "or", etc., combined with punctuation marks, placeholders for named entities, and regular expression elements. For example, the pattern "such NP as {NP ,}* {(or | and)} NP" can be applied to examples like "...works by such authors as Herrick, Goldsmith, and Shakespeare", and extract the following relations: is-a("author", "Herrick"), is-a("author", "Goldsmith"), is-a("author", "Shakespeare") [6,10]. The main advantage of rule development is that, due to their declarative nature, these patterns are understandable by humans and the effects of change are directly visible when compared to a machine learning model, which requires a training phase and an extraction phase. In general, the quality of the information extracted is quite high, but recall is usually low [11]. This approach has as its main issues scalability - given the high cost of developing rules -, and managing large sets of rules.

## 3 Methodology

The strategy we adopted for extracting information is the one identified by the work developed by Marti Hearst [5], where information is extracted with help of a set of lexical-syntactic patterns. This choice is due, among other factors, to the predictability of the DHBB texts, whose writing follows a fairly standardized structure. Furthermore, the AC/DC workflow allows us to improve the identification of certain types of information in the corpus through the creation of rules and lexicons of different semantic fields. Thus, new annotations are progressively incorporated into the corpus, to be accessed in search expressions.

The parser used was PALAVRAS, chosen for a number of reasons. The main ones are that this analyzer is specially designed for the Portuguese language, being considered one of the best within the chosen approach, with good syntactic and semantic analysis quality, and is also adopted by Linguateca for processing all corpora included in the AC/DC project [12]. PALAVRAS is rule-based, following constraint grammar (CG) and also performs NER. The tagging of candidate named entities is made in three levels: i) known lexical entries and lists of additional terms; ii) pattern-based prediction (morphological module); and iii) context-based inference for unknown words. Furthermore, the parser joins fixed expressions with the non-compositional semantic-syntactic function for MWEs, creating composite tokens and facilitating token-based CG syntactic rules [3].

The corpus is available through AC/DC, a service developed by Linguateca to make annotated corpora accessible through a web interface. The system employs the

IMS Open Corpus Workbench (CWB), a collection of open source tools aimed at questioning corpora enriched with linguistic annotation. CWB sees the corpus as an entity with its own integrity that can be interrogated but never changed, allowing for several different levels of annotation [14]. By being included in the AC/DC environment, queries to the DHBB – expressed through the CQP (Corpus Query Processor) language – can return different types of information, such as concordances or distributions, using extended regular expressions over linguistic annotation and other information present in the corpora.

A predetermined set of semantic tags assigned by PALAVRAS was used to identify some classes of entities in the corpus that we were interested on, namely <hum> for person, <org> and <inst> for organization/institution, <party> for political party, <occ> and <event> for event, <civ> for place, <tit> for document/work and <hprof> for professional role. A class able to bring together instances related to government plans, programs, agreements, treaties, laws, decrees, codes and all sorts of political formulations was missing, so we created a new class for this purpose: <titfpol>. The annotation of this class was made from a list containing instances that belong to it.

After the analysis made by PALAVRAS and other common processing by the AC/DC project, the structure of each entry is now in a pseudo-xml format, with segmented paragraphs and sentences, embedded metadata, and syntactic and semantic tags assigned to tokens. The result is a corpus containing annotation of words, lemmas, grammatical categories (pos), verb tenses, syntactic function and additional semantic information.

To improve NER, we created lists of entities obtained from three sources: i) instances existing in categories on Wikipedia; i) instances found with the AntConc concordancer using lexical clues applied to the corpus; and iii) instances identified by PALAVRAS and manually reviewed. These lists comprise 25,970 organizations, 1,250 policy formulations, 18,488 persons, 350 events and 1,011 political parties.

An extra layer of semantic annotation was then added with help of the *corte-e-costura* (cut-and-sew) tool developed for this purpose [13]. In general terms, the process starts from an initial lexicon whose correspondences in the corpus are noted as belonging to the target semantic field. Through context analysis of the annotated words, specialization or elimination rules are created to correct ambiguous cases. Using the same tool, we tackle two problems that affect the identification of some entities in the corpus. The first is due to errors in the segmentation of proper names, as in the case of *Eugênia Lopes de Oliveira Prestes de Macedo Soares*, which was automatically recognized as being two proper names instead of one: *Eugênia Lopes de Oliveira Prestes* and *Macedo Soares*. Adding some manual rules in the *corte-e-costura* we were able to join the two segments. The second problem is related to the various forms of writing the names of the biographees. It was necessary to both identify these variations in the corpus (lemmas) and their posterior unification. So we retrieved all proper names that PALAVRAS annotated as being “human” and that were not identical to any entry name, checked manually this list, and created a correspondence table indicating which lemmas corresponded to which biographees (called “grounding” in [9]). This is an iterative process: the list of lemmas is obtained, the correspondences are made; we get a new list, new matches, and so on.

After these preprocessing tasks, whose goal is to perfect the annotation in the corpus, we proceed to the IE stage proper, where we identify a set of patterns that can be tested and evaluated regarding their productivity in relation to the DHBB. We can summarize this new process as follows: for each theme, we observe in a sample of entries how the sentences that bring the desired information are constructed and we separate as many of them as possible, considering diversity and scope. We then translate these constructions into lexical-syntactic patterns with regular expressions, iteratively testing them until we get the sentences we are interested in. Finally, we concatenate all expressions, query the corpus, and postprocess the results using R. The excerpts of interest are isolated and synthesized, and the specific information is extracted with finer rules, to be then crossed with metadata.

## 4 Extraction evaluation

The first information to be extracted was the year of birth of the biographees. Only one pattern was needed, because the context in which this information appears in the entries follows a certain pattern: it is always located in the first paragraph, and is preceded by the character's name, with few variations. For example: “Moroni Bing Torgan was born in Porto Alegre on June 10, 1956.” or “Álvaro Francisco de Sousa was born on February 28, 1903.” The pattern created was:

[classe="bio.*" & dicionario="dhbb" & pos="PROP.*"]+ [: pos!="PROP.*" :][]{0,1} [lema="nascer" & word!="nascido|nascer"] [pos="PRP.*"][]{0,21} [pos="NUM.*|ADJ.*"] [word="de"]? [pos="N.*"]? [word="de"]? [pos="NUM.*"]? [pos="PU"].

Basically, it means: “if the DHBB entry is biographical, find a compound proper name, followed by the lemma 'born' and a preposition. Keep going until find the first number or modifier of the sentence (day, year or month) and continue to the next comma or period; in this range there may or may not be another number (possibly the other part of the date). This pattern brought 6,455 occurrences, which were checked manually. Overall, the rate of correct answers was around 98%, considering the entries that were not retrieved due to the absence of this information in the text (true-negative ones).

The same procedure was applied to extract information about the education background of the biographees and their family ties. In the first case, 11 patterns retrieved 10,565 excerpts from 5,627 entries, which represent around 83% of the total DHBB biographical entries. Each of the occurrences was manually checked to identify cases in which the lexical-syntactic construction met one of the defined patterns, but the information contained therein was not valid. If the extracted sentence mentions attended courses, obtained titles, admission to universities and related events, then it is valid, otherwise it is not. It is not possible to know how many and which occurrences were not recovered throughout the corpus, so the assessment does not include a measure for recall. But we considered that 99.1% precision was quite promising.

In the case of family relationships, we searched for patterns in a selection of ten entries whose holders are already known to have family ties with other politicians. The first step was to look for sentences where family ties occur. Looking closely at each of them, it was possible to classify them into valid and non-valid relationships,

based on both a supposed family relationship of the politician with other people, and relationships of third parties mentioned in the entries. Relationships that were considered not valid are cases in which the family term does not represent a direct link with anyone or does not demonstrate any family relationship, as in "...feeding pregnant women, young *mothers* and children". In addition to a perfect identification among biographees that uncovered 35 cases already grounded, we created rules to identify other family relations. We created thus 33 patterns that retrieved 6,220 such relationships in the corpus. Manual verification[3] was restricted to 198 random cases, yielding an average precision of .59. It is important to say that our interest here was not to know whether family ties were correctly identified using these patterns (and most of them were), but rather to measure the proportion of family relationships among politicians that can be found in the corpus. Furthermore, not all biographies can be considered politicians. In our understanding, a politician is someone who is invested in his position through election, nomination or designation, usually members of the executive and legislative branches. Positions that serve merely for bureaucratic jobs, such as technical advisers and consultants, whether executive, legislative, judiciary branches or military, are generally not considered politicians, although they are involved in government decision-making processes. Ordinary citizens such as activists and civic leaders are not considered politicians, even though they may be public opinion makers.

Summing up, to extract the year of birth of the biographees, the F-measure was .99 (their date of their first position is included in the metadata), to extract family relationships among politicians, the (estimated) precision was .59 and for information on education training, the precision reached .991.

These extractions, in turn, allow us to make a distant reading of DHBB that shows i) a drop in the average age in the entry of politicians into the public career, who start to position themselves more and more under 40 years of age, mainly those born from the 1960s; ii) a sharp decline in military training, especially for the post-1920 generations, showing that civilian training replaced military training as the preferred path to reach important political positions; and iii) family ties in politics as a phenomenon that remains over time at very significant rates, often representing more than half of the members of certain categories.

## 5 Distant reading DHBB

With the data obtained in the extraction, we can carry out some distant reading about the domain of the corpus. The distribution of ages ranging from the generation born before 1920 to the beginning of the 2000s showed us that those aspiring to political careers are getting in the public service steadily younger. The difference in the average between the so-called generation 1 (born before 1900) and generation 6 (born after the 1980s) falls almost by half, that is, if it was more common to start a public career around the age of 50 years old hundred years ago, over time this changed until it reached an average of 27 years old in the last generation (see Fig. 1).

[3] Which can be inspected at: https://www.linguateca.pt/acesso/dhbb/verifFamiliaDez2021.html

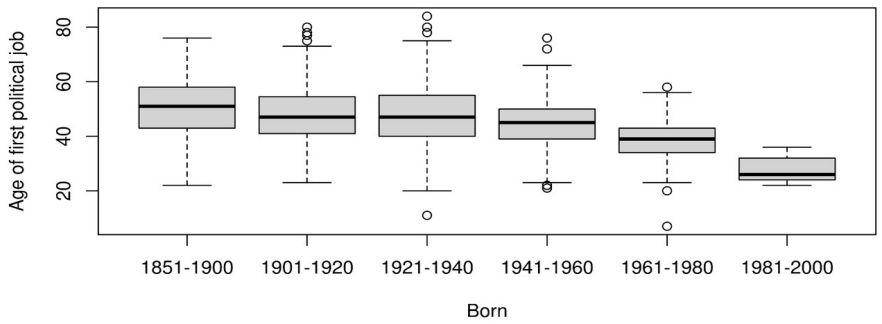


**Fig. 1.** The age of starting public careers, per generation

Regarding the education background of the politicians, among the 48 found areas, in the overall reckoning law school appears preponderant in all generations, followed by military training, with just one third of the first. Degrees in engineering, medicine, administration and economics come next, the latter two being practically tied. Across times, we found that the most significant transformation was the decline in military training from the second (born between 1901 and 1920) to the third generation (born between 1921 and 1940), decreasing almost by half. In fact, the lower presence of politicians with military training from the third generation onwards suggests that civilian training replaced the military career as the most suitable way to reach important political positions.

About family ties in politics, although it is not possible to determine how deeply entrenched Brazilian politics is with regard to family dynasties, since the logic of domination by kinship occurs also in other branches that the DHBB does not cover (like states and municipalities, Executives and Legislative branches), it is possible to see that it is a phenomenon that persists over time, at very significant rates. Presidents and senators are the politicians who most appear with family ties, 50% and 35% respectively, this being most perceived in the first generations (in the present version of DHBB). Ministers and deputies follow with 18%, keeping a stable average over generations. Unfortunately, most of the politicians elected in 2018 have not yet been included in DHBB, and therefore there is no way to study the current context, which would be quite interesting.

## 6 Final considerations

Our aim in this paper was to investigate the possibility of extracting useful, diverse and high-quality information from a corpus of encyclopedic text. We explored some approaches to automatic information extraction and described a methodology based on the use of textual patterns, where extraction rules are manually created and rely mainly on lexical-syntactic aspects and semantic annotation of the corpus. We tested

the proposed approach in three cases. In general, the results showed high accuracy in the quality of the information extracted.

One big challenge we became aware of during this investigation was how to find a balance between a sufficient number of patterns and a good enough coverage. This is about the quantity – and consequently, the manual work required – of expressions that should be created and applied, which is not easy to predict because of the language's own natural expressiveness.

Although the methodology itself is not new, this study explored issues that have been addressed in debates where the humanities disciplines have been modified with the use of computational techniques, in order to assess the opportunities that open up in this potentially innovative scenario. In our view, these new tools can indeed lead to the expansion of academic research along various alleys, both in terms of methods renewal and knowledge production. Certainly difficulties exist and the challenges are many, but the possibilities open by innovation scenarios lead to a continuous and deserved effort to try to overcome them, if we invest enough work in tailoring the material to our research needs.

The main contributions of this work are: the creation of an annotated encyclopedic corpus made available for language and humanities studies; the presentation of a methodology based on a philosophy of cyclical enrichment: the more information is obtained, the more it is added to the corpus itself; and the compilation of a set of patterns that can be adapted to other corpora containing a similar type of annotations [7].

## References

1. Abreu, A., Lattman-Weltman, F., de Paula, C.J. (eds.): Dicionário Histórico-Biográfco Brasileiro pós-1930. CPDOC/FGV, Rio de Janeiro, Brazil, 3 ed. (2010).
2. Agt-Rickauer, H.: Supporting Domain Modeling with Automated Knowledge Acquisition and Modeling Recommendations (Thesis). Elektrotechnik und Informatik der Technischen Universitat Berlin, Germany (2019).
3. Bick, E.: Functional Aspects on Portuguese NER. In Santos, D., Cardoso, N., ed., Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. Linguateca, pp. 145-155 (2007).
4. Golshan, P. N., Dashti, H. R., Azizi, S., Safari, L.: A Study of Recent Contributions on Information Extraction. arXiv preprint :1803.05667 (2018).
5. Hearst, M.: Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th conference on Computational linguistics, v. 2, Coling'92, pp. 539–545, Stroudsburg, PA, USA (1992).
6. Hearst, M.: Automated discovery of WordNet relations. In Fellbaum, C., ed., WordNet: An Electronic Lexical Database, p. 131-151. MIT Press, May (1998).
7. Higuchi, S.: Extração automática de informações: uma leitura distante do Dicionário Histórico-Biográfico Brasileiro (DHBB). Thesis. PUC-Rio. Rio de Janeiro, Brazil (2021).
8. Higuchi, S., Freitas, C., Cuconato, B., Rademaker, A.: Text Mining for History: First Steps on Building a Large Dataset. In Calzolari, N. et al., ed, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan (2018).

9. Higuchi, S., Freitas, C., Santos, D.: Distant reading Brazilian politics. In Proceedings of 4th Conference of The Association Digital Humanities in the Nordic Countries, pp. 190-200. Copenhagen, Denmark (2019).
10. Jurafsky, D., Martin, J. H.: Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River, USA: Pearson/Prentice Hall (2009).
11. Makarov, P.: Automated Acquisition of Patterns for Coding Political Event Data: Two Case Studies. In Proceedings of Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pp. 103–112. Santa Fe, New Mexico, USA (2018).
12. Santos, D., Bick, E.: Providing Internet access to Portuguese corpora: the AC/DC project. In Gavrilidou, M., Carayannis, G., Markantonatou, S., Piperidis, S., Stainhauer, G., ed., Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000), pp. 205-210. Athens, Greece (2000).
13. Santos, D., Mota, C.: Experiments in human-computer cooperation for the semantic annotation of Portuguese corpora. In Calzolari et al., ed., Proceedings of LREC 2010, pp. 1437-1444. Valetta, Malta (2010).
14. Santos, D., Ranchhod, E.: Ambientes de processamento de corpora em português: Comparação entre dois sistemas. In Rodrigues, I., Quaresma, P., ed., Actas do IV Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR'99), pp. 257-268. Évora, Portugal (1999).