Distant reading Brazilian politics

Suemi Higuchi[1,4], Diana Santos[2,3], Cláudia Freitas[4,2], and Alexandre Rademaker[5,6]

[1] CPDOC/Fundação Getulio Vargas, Praia de Botafogo, 190, Rio de Janeiro - Brazil
[2] Linguateca http://www.linguateca.pt
[3] University of Oslo, HF, ILOS, Pb 1013 Blindern, Oslo, Norway
[4] PUC-Rio, Rua Marquês de São Vicente, 225, Gávea, Rio de Janeiro - Brazil
[5] IBM Research, Avenida Pasteur, 138, Urca, Rio de Janeiro - Brazil
[6] EMAp/Fundação Getulio Vargas, Praia de Botafogo, 190, Rio de Janeiro - Brazil
suemi.higuchi@fgv.br, d.s.m.santos@ilos.uio.no,
claudiafreitas@puc-rio.br, alexrad@br.ibm.com

Abstract. In this paper we propose the use of digital humanities tools to "read" and obtain aggregated information on Brazilian politics. After presenting briefly the resource and its annotation, we describe the kinds of searches already possible, our work for grounding human entities, and some results on family relationships among Brazilian politicians.

Keywords: information extraction · Portuguese · Brazilian history

1 Introduction

The intricate relationship between traditional practices of recording knowledge and new technologies is the indelible mark of the Digital Humanities (DH) movement. They incorporate the methods and issues developed by the human and social sciences, while mobilizing the unique tools and perspectives opened by digital technology [6]. In the area most closely linked to language and literature, where there are millions of digital collections to study, observations made at a distance and from various perspectives are only possible with the aid of computers and statistical techniques capable of reducing the literature to a set of interesting and manipulable data. In this sense, work with annotated corpora in order to automate (and therefore eventually obtain) information like characters, plot, or events, is becoming mainstream. In this paper, we describe some work in this vein, concerning a Brazilian resource named *Dicionário Histórico-Biográfico Brasileiro* (DHBB for short). Although coined as "dictionary", DHBB has an encyclopaedic format, with long entries written by experts, describing relevant actors in Brazilian history. It is a reference work and, as such, it is not intended to be read in a linear (or conventional) way, but to be consulted instead. Within the scope of Digital Humanities, with its tools, methods and resources, to get the vast amount of information spread among DHBB pages in a structured way is a challenge as desirable as predictable. The following sections present the strategies and results obtained so far.

2 The Dicionário Histórico-Biográfico Brasileiro

DHBB is an encyclopedia developed and curated by Centro de Pesquisa e Documentação de História Contemporânea do Brasil, from Fundação Getulio Vargas (FGV), and is an important resource for all research, nationally and internationally, interested in Brazilian politics [1]. It contains information ranging from the life trajectory, education and career of the individuals, to the relationships built between the characters and events that the country has hosted.

DHBB was first published on paper in 1984, in four volumes containing 4,500 entries. In the 2001 update, the resource was increased by one more volume reaching a total of 6,620 entries, and in 2010 its material was made available on the internet, with about 7,500 entries. Currently the DHBB holds ca. 8,000 entries and is continually updated and improved[7]. The information system has the following structure: per entry, its designation, the kind of entry (biographical or thematic), and the text in a text field. 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 [9] and [10]. Each entry became a single file that received a unique identifier, and new metadata were added, such as the gender of the biographed and the political role s/he had. All text files are available in github.[8]

Later on we converted DHBB into an annotated corpus, subject to syntactical analysis by PALAVRAS [2] and semantic annotation by AC/DC [5], and made available through Linguateca's site[9].The DHBB resource was thus enriched with syntactic and semantic information, quite useful for doing historical research.

2.1 General characterization of the corpus

In this section we give some figures about DHBB's content. Some of it comes directly from the metadata associated with the previous versions, other cases are a direct consequence of being in an annotated form. As we are still in a preliminary phase of work, it is possible that some of these numbers will change with time, but we believe they are already good indicators of the richness of the material.

The universe we are working on from the DHBB is ca. 7,600 entries corresponding to 237,561 sentences, 7 million tokens and 97 thousand lemmas. 1.3 million tokens correspond to proper names, almost 100,000 different ones. Of those, roughly 41,000 have been analyzed as person names, 36,000 as organization names and 4,000 as places names by PALAVRAS. There are 5,953 biographical entries, the rest being thematic. Of the biographical ones, the vast majority refers to men (5,780), only 184 concern women. In Table 1 we present an overview of the roles present (obviously, the same person can have more than one role throughout her life).

[7] Oficial webpage: `https://cpdoc.fgv.br/acervo/dhbb`
[8] Available at `https://github.com/cpdoc/dhbb`.
[9] Available at `https://www.linguateca.pt/acesso/corpus.php?corpus=DHBB`

Table 1. Description of DHBB in terms of political roles.

Role or job	occurrences
Presidentes do Brasil (presidents of Brazil)	26
Ministros (ministers)	776
Ministros do STF (judges of the highest court)	96
Ministros do STM (judges of the highest military court)	118
Senadores (members of the Senate)	627
Deputados Federais (parliament members)	3,835
Militares (Army officers)	704
Participantes de revoluções (revolution participants)	368
Jornalistas (journalists)	196

2.2 A rich source of information

In the late 1980s, a study conducted by Michael Conniff [3] and [4] with a sample of 7% of the entries (about 250 biographies at the time), enabled him to locate important changes concerning age, education, social class and geographical origin in the Brazilian political elite by close reading all these entries.

By extracting manually the information he was after, he was able to map several interesting features of this elite. For example, in the beginning of the twentieth century, most Executive members were middle-aged or older men, who typically entered political life as second career, after having had other jobs. Later on, those who aim for a political career get increasingly younger. On average those born before 1900 start at 55, those born between 1901 and 1920 start at 37, and the ones born after 1921 start at 32 years old. As to formal education, the most common one is Law (44%) followed by military education (32%). Engineers and doctors follow with 12% and 5% each. The most definite change spotted by Conniff is the decline in military careers of politicians: while for those born before 1920, 37% had military education, for the ones born after 1920 only 10% had. and using SPSS[10] Until now, if a researcher is interested in e.g. the question of 'how did military politicians enter politics in Brazil, through revolution or legally?' s/he has to read every relevant entry. The same happens for the questions 'what is the path most frequently followed to attain the presidency?' or 'where do the highest military judges (*ministros do Superior Tribunal Militar*, in Portuguese) come from in terms of regions/states in Brazil after 1965?' or even 'what is the average age for a judge to enter the Supreme Federal Court?'

By annotating the free text with morphosyntactic information and several semantic domains, we hope to be able to get most of this information automatically. In DH terms, one could describe this as distant reading [7] for history.

[10] SPSS (Statistical Package for the Social Sciences) is a software package used for statistical analysis. See `http://www.ibm.com/software/analytics/spss/`

3 Enhancing the DHBB with further relevant information

In addition to the usual information in an AC/DC annotated corpus, we concentrated on named entity recognition. In particular, for this resource, the recognition of person names, places, organizations and political roles. Most of this is already provided by PALAVRAS, and we just checked whether there were systematic problems that should be corrected. (For example, names like *Eugênia Lopes de Oliveira Prestes de Macedo Soares* have been wrongly tokenized as two proper names instead of one – *Eugênia Loopes de Oliveira Prestes* and *Macedo Soares* –, but this is easy to correct with our rule-based tools for corpus annotation revision, described in [11]).

In addition, and due to the fact that the same politician can be referred to in several ways, especially in a context where s/he has been named before, we decided to do entity grounding: we want to assign to each person name the entity identifier it refers to, using as unique 'identifiers' the entry labels (see section 3.1 below). Also, we added information relative to family relationships to this corpus, as yet another relevant type of semantic information. We detail the processing done in the next subsections.

3.1 The grounding process

There are many more cases of distinct proper names than distinct human entities, and we want to identify who is who (i.e., to which entity they refer). So we created an attribute `id` that contains the entry identifier which describes that person in DHBB, and we try to assign it to all proper names which do have a "definition" in DHBB.

Table 2. Examples of correspondences created manually for the most frequent non-grounded proper names.

AC/DC lemma	Full name as entry in DHBB (id)
Aécio=Neves	Aécio Neves da Cunha
Alencar=Castelo=Branco	Humberto de Alencar Castelo Branco
Anthony=Garotinho	Anthony William Matheus de Oliveira
Getúlio=Vargas	Getúlio Dornelles Vargas
Lula	Luis Inácio da Silva

So our task is to annotate the 41,856 different human proper names (corresponding to 404,245 words) in the texts so that, if they refer to someone defined in DHBB, they receive the corresponding `id`. Of course, there is a lot of people (spouses, parents, etc.) which are mentioned in a biographical entry but are not necessarily politicians with a DHBB entry. In cases where such people have to be mentioned in rules (see below), they are assigned the label `NV`, which stands for "não verbetado" (not included as an entry). If some people are very often

mentioned in the DHBB but have not an entry of their own, they may be good candidates for future inclusion.

The semi-automatic grounding process is as follows. First, we annotated those proper names which are exactly equal to the entry form (usually the full name). This allowed us to ground at once 89,937 words. Then, we produced a (first) list of 116 abbreviations in the form illustrated in Table 2, and managed to increase the number of grounded proper names to 147,085. In a second iteration, adding 71 new correspondences (or abbreviation rules), we obtained 166,059 cases.

Another problem concerning proper names is that they can refer to different people, as Table 3 shows.

Table 3. Proper names of people including the word *Vargas* (excluding therefore organizations like *Fundação=Getulio=Vargas*).

Proper name	ocurrences
Vargas	3609
Getúlio=Vargas	1735
Ivete=Vargas	96
Benjamim=Vargas	52
André=Vargas	33
Lutero=Vargas	27
Alzira=Vargas=do=Amaral=Peixoto	18
Jorge=Vargas	9
Manuel=do=Nascimento=Vargas	7
Israel=Vargas	7
Manuel=Vargas	7
Darci=Vargas	6
Viriato=Vargas	6
Alzira=Vargas	5
Protásio=Vargas	4
Getúlio=Dornelles=Vargas	4

We could explore the following heuristic for ambiguous terms: mostly a shorter form will refer to the entry subject. In the case of the string *Vargas*, when alone and not in entries also including the name *Vargas*, it almost always refers to Getúlio Vargas, a very influential Brazilian president (in 1834-1945 and 1951-1954). Nevertheless, this is not always the full story, and we implemented a specific form of abbreviation rules which includes exceptions, as displayed in table 4.

Finally, another task that we foresee is doing (easy) anaphoric reference resolution by taking into consideration the person who is being biographed. In the following examples, the underlined proper names refer to the main entry, in bold.

> **Getúlio Dornelles Vargas** nasceu em São Borja (RS) no dia 19 de abril de 1882, filho de Manuel do Nascimento Vargas e de Cândida Dornelles

Table 4. Cases where the shortest form of the name corresponds to the entry name and cases where it does not.

AC/DC lemma	entry where it occurs	correspondence entry name (id)
Vargas	Getúlio Dornelles Vargas	Getúlio Dornelles Vargas
Vargas	José Israel Vargas	José Israel Vargas
Vargas	Alzira Vargas do Amaral Peixoto	Getúlio Dornelles Vargas
Vargas	Benjamim Dornelles Vargas	Getúlio Dornelles Vargas
Vargas	Lutero Sarmanho Vargas	Getúlio Dornelles Vargas

> Vargas. Vargas era descendente de uma família politicamente proeminente em São Borja, região de fronteira com a Argentina, palco de rumorosas lutas no século XIX. O pai de Getúlio, Manuel do Nascimento Vargas, combateu na Guerra do Paraguai, distinguindo-se como herói militar.

3.2 Family relationships

One semantic domain that we are especially interested in can be illustrated by the generic question 'How many politicians in the last decades belong to a family of politicians?' In Brazil there are powerful families since the colonial period which can be said to form political dynasties. By pushing their children and relatives to the parliament and the senate, they have been analysed [12], [8] as strong power-maintaining devices. Has this phenomenon increased, or decreased, lately? Does this practice only concern rich families of the periphery, or has it also pervaded other less traditional groups? We know this information is diluted in the thousands of DHBB entries, and we have started to add semantic annotation on family relations in order to deal with it.

In AC/DC there are currently several domains that have been subject to thorough annotation (colour, body, emotions, health, clothing), and for DHBB we added family. We created a list of family-denoting words which were integrated in the semantic annotation process, and we are currently creating rules (following the explanation in [11] to improve and correct the annotation. The lists include 50 family-denoting nouns, 10 family-related verbs and 9 other family-related terms so far.

Even though this is in a preliminary stage, Table 5 shows the most common family relationships in DHBB, while Figure 1 shows in context several cases of family relationships among grounded politicians, using a simple search command.

4 Some distant reading

By concatenating in a single query the political role conveyed in the metadata, simple lexicosyntactic patterns, and semantic information, it is feasible to search for things such as: a) formal education of the members of the Chamber of Deputies (*deputados federais*) elected by a specific location - for example,

Table 5. Most frequent family ties in DHBB. The second translation refers to the possible meaning of the plural. Eg. *filhos* can mean children (sons and daughters), and *irmãos* siblings.

Lemma	occurrences
filho (son, child)	9444
pai (father, parent)	1488
irmão (brother, sibling)	1342
filha (daughter)	1144
mulher (wife)	523
tio (uncle)	312
primo (cousin)	287
esposa (wife)	248
mãe (mother)	230
sobrinho (nephew)	186
parente (relative)	172
irmã (sister)	131
marido (husband)	130
avô (grandfather)	116
cunhado (brother in law, in-law)	102

ıar=9: Devido à disputa pelo governo do estado de Mato Grosso entre Mário Correia da Costa e **Fenelon Müller, irmão de Filinto Müller**, chefe de polícia do Distrito Federal, ocorreu um impasse político naquela unidade da Federação .

ıar=8: Tal movimento teria sido sustado graças a uma enérgica intervenção do ministro do Exército, o general **Orlando Geisel, irmão de Ernesto Geisel** .

ıar=18: Em setembro, depois de reconciliar-se politicamente com as correntes de Luís Viana Filho e Antônio Lomanto Júnior, foi eleito governador pelo colégio ıleitoral do estado, tendo como vice-governador **Luís Viana Neto, filho de Luís Viana Filho** .

ıar=1: Seu sobrinho, **José Brás Pereira Gomes, filho de Venceslau Brás**, foi deputado federal por Minas Gerais entre 1933 e 1937 e constituinte em 1934 .

ıar=10: No Rio Grande do Sul, o governador **Leonel Brizola, cunhado de João Goulart** e fiel à Constituição, organizou um movimento de resistência à ıposição militar lançando a campanha da legalidade, com o objetivo de assegurar a posse do vice-presidente .

Fig. 1. Getting family relations among grounded entities in AC/DC.

the state of Rio de Janeiro; or b) their birthplaces. The results show that we have so far in DHBB 333 politicians who held the position of deputy by Rio de Janeiro at least once, were born in 117 different cities and their most common education background is: law (65), engineering (15), medicine (11), economics (7) and business school (5), followed by theology (4) and geography(4). When we contrast these results with the education background of all brazililian federal deputies, the picture remains almost the same.Figure 2 shows that theology is a common background, and this is similar to the situation in all states.

5 Future work

One of the goals of presenting this resource to a DH community is to get input as to further developments and intelligent ways of reading it distantly. We plan to extract all sorts of information from DHBB and crosscheck the data with small probes done by close reading.

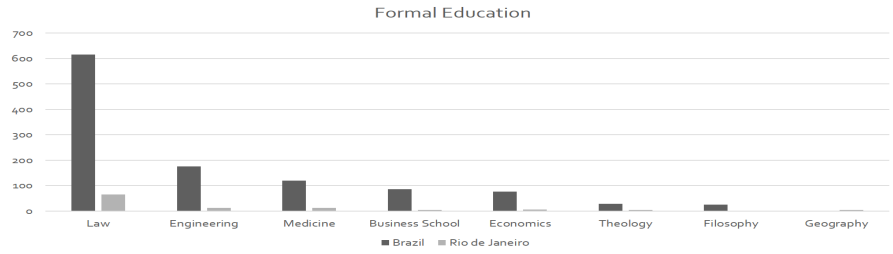

Fig. 2. Formal education of Brazilian federal deputies

We plan to annotate other semantic domains that appear relevant to studies of Brazilian politics and that are brought to light by the users, things like political parties, governments and alliances. And, in a longer perspective, we also envisage map-based and chronological visualization capabilities, to endow DHBB users with different ways of interacting, and comprehending the data.

References

1. de Abreu, A.A., Lattman-Weltman, F., de Paula, C.J. (eds.): Dicionário Histórico-Biográfico Brasileiro pos-1930. CPDOC/FGV, Rio de Janeiro, 3 edn. (2010), available at `http://cpdoc.fgv.br/acervo/dhbb`
2. Bick, E.: The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press (2000)
3. Conniff, M.: O DHBB e os brasilianistas. In: FGV, E. (ed.) CPDOC 30 Anos. Editora FGV/CPDOC, Rio de Janeiro (2003)
4. CONNIFF, M.L.: A elite nacional. Por outra história das elites. Rio de Janeiro: FGV pp. 99–121 (2006)
5. Costa, L., Santos, D., Rocha, P.A.: Estudando o português tal como é usado: o serviço ac/dc. In: Proc. of STIL 2009) (2009)
6. Dacos, M.: Manifeste des digital humanities (2011)
7. Moretti, F.: Conjectures on world literature. New left review pp. 54–68 (2000)
8. de Oliveira, R.C., Goulart, M.H.H.S., Vanali, A.C., Monteiro, J.M.: Família, parentesco, instituições e poder no brasil: retomada e atualização de uma agenda de pesquisa. Revista Brasileira de Sociologia-RBS **5**(11) (2017)
9. 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 (2014)
10. 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)
11. Santos, D., Mota, C.: Experiments in human-computer cooperation for the semantic annotation of portuguese corpora. In: Calzolari et al (eds) Proceedings of LREC 2010. European Language Resources Association (2010)
12. Schoenster, L.: Clãs políticos seguem dominando congresso na próxima legislatura. Transparência Brasil. Disponível em http://www. excelencias. org. br/docs/parentes_ pp. 202015–2018 (2014)