Exploratory Information Extraction from a Historical Dictionary

Valeria de Paiva[‡], Dário A. B. Oliveira[†], Suemi Higuchi[†], Alexandre Rademaker[*] and Gerard de Melo[§]
[‡] Nuance Communications, USA, valeria.depaiva@nuance.com
[†] FGV/CPDOC, Brazil, {dario.oliveira,suemi.higuchi}@fgv.br
[*] FGV/EMAp and IBM Research, Brazil, alexrad@br.ibm.com
[§] Tsinghua University, Beijing, China, gdm@demelo.org

Abstract—**We describe a preliminary project of extracting information from an extant dictionary of historical biographies, the "Dicionário Histórico-Biográfico Brasileiro" (the Brazilian Historical and Biographical Dictionary, shortened as DHBB), a long-standing project at the 'Centro de Pesquisa e Documentação de História Contemporânea do Brasil' (CPDOC) of the 'Fundação Getulio Vargas' (FGV). For information extraction, we rely on Natural Language Processing tools such as FreeLing as well as our own resources NomLex-PT, a lexicon of nominalizations, and OpenWN-PT, a Portuguese version of Princeton's WordNet database. While our project currently highlights the potential of information extraction in a fun exploratory manner, we also discuss the engaging of historians interested in the affordances of digital tools.**

I. Introduction

The ever-increasing availability of digital tools and vast amounts of data are more than just a useful aid for scholars in the humanities. It is now quite evident that they in fact enable entirely novel forms of research that were not feasible in the past.

In this paper, we focus on exploratory analyses of a Brazilian dictionary of historical biographies, the "Dicionário Histórico-Biográfico Brasileiro"[1], Brazilian Historical and Biographical Dictionary or DHBB for short. Over the course of many years, the DHBB has been created and maintained at the Centro de Pesquisa e Documentação de História Contemporânea do Brasil (CPDOC) of the Fundação Getulio Vargas (FGV) in Rio de Janeiro, Brazil.

Our analyses rely on tools and resources from Natural Language Processing (NLP). In particular, we rely on FreeLing, a suit of multilingual NLP components, as well as on our resources NomLex-PT, a lexicon of Portuguese language nominalizations, and OpenWN-PT, a Portuguese version of Princeton's semantic lexicon WordNet [2]. These tools enable us to extract semantic information from the dictionary entries, going far beyond standard keyword search interfaces that are typically used to browse such cultural heritage resources.

II. The DHBB Dictionary

The Brazilian Historical-Biographical Dictionary (Dicionário Histórico Biográfico Brasileiro or DHBB in Portuguese) was conceived with the basic purpose of providing researchers and scholars with organized and systematic information about personalities and themes considered noteworthy in the recent history of Brazil.

The time frame covered by the DHBB encompasses the historical period that began with the "Revolution of 1930", a period in Brazilian history marked as involving rupture and significant renewal in the political elite, as well as enabling new institutions and movements. There are about 7,500 biographic or thematic entries in the DHBB, covering people, institutions, organizations, and events. The vast majority of entries are biographical in nature, with over 6,500 biographies against some 1,000 thematic entries.

The project was initiated and pursued by the CPDOC, which has as its main goal the preservation of relevant document sources of the country's history and the development of research tools and methods for historical and economical research on the Brazilian cultural heritage. The institution's document collection serves as a privileged reference source for such research.

The first edition of the DHBB dates from 1984, and since then two major updated editions have appeared. The first of these was in 2001, while the most recent major update in 2010 was innovative in that it brought full availability of the contents of the DHBB to the Internet. In 2014, a team of editors at CPDOC was formed to undertake a new update, from which we expect the addition of numerous new entries as well as updates of the existing ones, since significant numbers of them describe people that are still alive and active.

The DHBB is a benchmark of scholarly work, providing in a concise and unified way, a significant amount of data which had been dispersed in a number of primary and secondary sources beforehand. The project aims at objective and unbiased entries, avoiding as much as possible any ideological or personal judgments. The CPDOC researchers carefully revise all entries to ensure accuracy of the information and to ensure a uniform style across different entries.

A. Characteristics of the DHBB

Since the first version of DHBB, the editors have tried to standardize the different types of information included in the dictionary. For this, they developed general writing guidelines that state how the information should be written, the preferred order of stating facts, and so on. For instance, there are rules for writing names of people, institutions, political parties, social movements, treaties, historical episodes, and places. Some of these rules aimed at facilitating information retrieval [3], [4] in the earlier printed versions of the DHBB or at making the dictionary accessible to the general public. For example, the spelling of proper names follows some general orthography principles in Brazilian Portuguese, which means that they can be modified from the original. The letters ‘Y’ and ‘W’ are replaced by ‘I’ and ‘V’ (every ‘Maya’ will be ‘Maia’, ‘Oswaldo’ will become ‘Osvaldo’), in some cases ‘Z’ becomes ‘S’ (then ‘Souza’ becomes ‘Sousa’ and ‘Menezes’ becomes ‘Meneses’). However any foreign names are kept in their original form. Such rules may appear unusual and dispensable in modern times when data is digitalized and expected to be retrieved by search engines capable of answering more advanced requests with wildcard, range, and fuzzy queries.

Other rules were developed to keep a uniform presentation of the text. For example, quotes are always used in citations and also in expressions of recurrent use in historiography such as names of specific historic episodes as “escândalo da mandioca” (manioc scandal), “pacote de abril” (April’s package), “Operação Cristal” (Crystal Operation). Names of holidays are always capitalized (Semana Santa/Holy Week, Dia do Trabalho/Labor Day), as are the names of monuments (Monumento aos Pracinhas/Monument to the Unknown Soldier, Memorial JK/Juscelino Kubischeck Memorial) or of established historical periods or episodes (Revolução de 1930/Revolution of 1930, Revolta Comunista de 1935/Communist Insurgency of 1935, República Velha/Old Republic), etc.

Of primary concern to the editors of the DHBB are the inclusion criteria for adding new entries about the political history of the country and the question of which perspective should guide the selection of entries. For biographies, anyone who has occupied a relevant position at the federal level of the country’s public administration was included, although this means that local and municipal levels of administration have been disregarded. An important sample of Brazil’s civil society is represented in the dictionary, with the inclusion of certain presidents of organizations, entities, and private companies. The main leaders of rebellions and protagonists that held informal positions of power were included as well.

With regards to thematic entries, the DHBB describes political parties and political movements, organizations, and historical events. Also included are constitutions, decrees, laws, and codes, certain current and basic concepts of political history, economic and administrative institutions, national impact newspapers, and topics pertaining to foreign relations issues.

B. Biographical Entries

The generic format of a biographical entry is as follows. Below the title of the entry, there is a summarizing description of the person, with key positions held and the respective periods the position was exercised. This header, which aims to facilitate identification of the biography, employs acronyms and abbreviations whose meaning can be found in a predefined list. Because the positions are what justifies the inclusion of the person in the DHBB in accordance with the inclusion criteria, this introductory element is called ‘justification’.

> GOULART, João
> * dep. fed. RS 1951 e 1952-1953; min. Trab. 1953-1954; dep. fed. RS 1954; vice-pres. Rep. 1956-1961; pres. Rep. 1961-1964.

The next paragraph of the entry always consists of the full name of the person, their place of birth (followed by the abbreviated state in brackets), date of birth, parents, and additional information pertaining to the person’s origin and birth. For the example above, we find the following. [1]

> *João Belchior Marques Goulart nasceu em São Borja (RS), no dia 1º de março de 1919, filho de Vicente Rodrigues Goulart e de Vicentina Marques Goulart. Desde criança recebeu o apelido de Jango, comum no sul do país.*
>
> João Belchior Marques Goulart was born in São Borja (RS), on the first of March, 1919, the son of Vicente Rodrigues Goulart and Vicentina Marques Goulart. He was known by his childhood nickname, “Jango”, a nickname common in the south of the country.

The rest of the textual body consists of the following sequence of elements: academic training, professional activity, political action, date of death, names of spouse and children, works written by and about the person, date of publication. The penultimate paragraph list the names of the authors of the entry and the last paragraph cites the references and sources consulted for the entry.

The entries go beyond simple reports on the life and career trajectories of the people described. The text also characterizes their relationships with other people, their leadership style, their strengths and weaknesses, and their legacy.

C. Thematic Entries

There is a fixed order to be obeyed in the writing of the thematic entries. After the main title, the introductory paragraph of the thematic entry always contains a summary of the topic, which includes: place and date of start/end of the episode, the names of the involved parties and a brief explanation of its main objective. One example would be:

> REVOLUÇÃO DE 1930
> *Movimento armado iniciado no dia 3 de outubro de 1930, sob a liderança civil de Getúlio Vargas e sob*

[1] The processing of the corpus, at the moment, simply moves the information from the justification part to the metadata of the entry. Later we hope to process this information in a semantic way.

a chefia militar do tenente-coronel Pedro Aurélio de Góis Monteiro, com o objetivo imediato de derrubar o governo de Washington Luís e impedir a posse de Júlio Prestes, eleito presidente da República em 1º de março anterior. O movimento tornou-se vitorioso em 24 de outubro e Vargas assumiu o cargo de presidente provisório a 3 de novembro do mesmo ano.

Armed movement which began on October 3rd, 1930, under the civilian leadership of Getúlio Vargas and under the military leadership of lieutenant colonel Pedro Aurelio de Gois Monteiro, with the immediate aim of overthrowing the government of Washington Luís and preventing Júlio Prestes, elected President of the Republic on March 1st, from taking office. The movement became victorious on October 24 and Vargas took over as interim president on November 3rd of the same year.

Subsequently, the rest of the textual body describes the background or origins of the episode or situation, as well as its structure, evolution, and dissolution. The last paragraph cites the bibliographic sources consulted.

III. Computational Linguistics Tools

We have been developing tools to process texts in (mostly Brazilian) Portuguese automatically. Our principal contribution so far is OpenWordNet-PT, an electronic dictionary and thesaurus automatically derived from Princeton's WordNet, but manually corrected by native speakers. We also build upon the excellent work of the FreeLing group, as described below.

A. FreeLing

FreeLing [5] is an open-source multilingual natural language processing toolkit, providing a wide range of analysis modules for several languages, including Portuguese. It offers text processing and language annotation facilities to NLP application developers, lowering the cost of building such applications. Quoting its main developer, Lluis Padró, 'FreeLing is customizable, extensible, and has a strong orientation to real-world applications in terms of speed and robustness. Developers can use the default linguistic resources (dictionaries, lexicons, grammars, etc), extend/adapt them to specific domains, or – since the library is open source – develop new ones for specific languages or special application needs.' We used FreeLing to lightly process the DHBB and to perform a few experiments with the contents of the result of this processing of the corpus.

B. OpenWordNet-PT

OpenWordNet-PT [6] is an electronic dictionary and thesaurus of Portuguese initially derived from Princeton's English WordNet and many other electronic dictionaries using automatic methods developed for the Universal WordNet project by de Melo and Weikum [7]. Parts of this data were then checked for correctness and extended with new glosses and lemmas by native speakers of Portuguese.

OpenWordNet-PT is a long-term project that has already proven useful to the community, despite being under current development. Francis Bond selected it as the wordnet for Portuguese in his Open Multilingual WordNet coordination [8]. Moreover, the project data is already used by the FreeLing community for some of its modules. Tables I and II summarize how OpenWN-PT has increased over the last two years. The number of synsets should be understood as the number of synsets from the Princeton Wordnet, with at least one Portuguese word included. Table II shows only the five classes more (and less) covered.

Table I
OpenWN-PT's coverage development

	2011	2013	increase
synsets	41,810	43,895	5%
words	52,220	54,125	3%
senses	68,285	74,054	8%

Table II
OpenWN-PT's synsets by classes

lexFile	openWN-PT	Princeton WordNet	percent
adj.ppl	5	60	8
verb.competition	100	459	22
noun.possession	271	1061	26
verb.creation	184	694	27
adv.all	979	3621	27
⋮	⋮	⋮	⋮
noun.phenomenon	324	641	51
noun.feeling	223	428	52
noun.object	908	1545	59
noun.location	2096	3209	65
noun.Tops	51	51	100

OpenWordNet-PT is available for download at https://github.com/arademaker/openWordnet-PT and can be consulted via Bond's Open Multilingual Wordnet interface at http://compling.hss.ntu.edu.sg/omw/cgi-bin/wn-grid.cgi.

C. NomLex-PT

We also developed a freely available computational lexicon for Portuguese, called NomLex-PT [9] that provides mappings between verbs and their nominalizations. This sort of lexicon is important, as computational systems typically lack information about connections between nouns and related verbs. Thus, a system encountering the noun *eleição/election* might not recognize its semantic relationship to the verb *eleger/to elect* and hence it may fail to recognize when different sentences or texts refer to the same events.

Some of our work on building lexical resources is manual and requires detailed expert analyses of linguistic data, which is known to be frustratingly time-consuming. At the same time, some of the resources that we would wish to provide already exist in other languages, and so a certain amount of translation can enable us to speed up the creation quite significantly. New cross-lingual induction algorithms are now mature enough to be applied to large crowd-sourced data collections like Wiktionary, which has grown significantly in recent years. Thus we used many of these freely available automatic and manually constructed resources, including the original English NomLex, and produced our own version for Portuguese. The resulting lexicon is available from https://github.com/arademaker/nomlex-pt.

IV. Named Entities in the DHBB

A dictionary of historical biographies is mostly about people, what they did and where and when they did it, so recognizing named entities (such as people, organizations, places, and times) in the text and correctly classifying their type is crucial to our task.

FreeLing offers an off-the-shelf named entities recognition (NER) module, which we used to make very preliminary experiments and to decide on further tools and methods. The program authors write that there are two different modules capable of performing NER. The basic module is simple and fast, and easy to adapt for use in new languages, provided that capitalization can serve as the primary clue for NE detection in the target language (which is the case for Portuguese). The estimated output quality of this module is about 85% (correctly recognized named entities). The second module relies on machine learning algorithms. It has a higher precision (over 90%), but is apparently slower than the basic module, and adaptation to new languages requires a training corpus plus some feature engineering.

We used the complete FreeLing stack of NLP modules, including the machine learning NER component, to process all the entries from the DHBB. In our analysis we considered only those named entities with at least 500 detected occurrences across the entire corpus.

Locations Regarding locations, we encountered 35 location names with more than 500 occurrences. Among these, there are two clear mistakes: "vargas" and "carta". Getúlio Vargas, former President and dictator of Brazil certainly has many places named after him, including a small town in the state of Rio Grande do Sul, and a main boulevard in Rio de Janeiro, named in his honour. However, this small town is presumably not significant enough in the context of locations in the history of Brazil to be mentioned at least 500 times. It is more likely that the influential former president has been misclassified as a location or that merely all the locations (roads, streets, squares) named after the former president are being identified. AS a second example, the word *carta/letter* (or Constitution when spelled with capital C) was apparently tagged by FreeLing as a place every time it appears in the pattern *da nova Carta*. Looking at the data and clustering by hand, we can see some of the problems that plague the work on named entities. Some locations can be meaningfully shortened (e.g. 'Rio' for 'Rio de Janeiro', while others cannot e.g. 'São Paulo'). Some names can be used for cities or states as is the case for the two main cities/states in Brazil (Rio de Janeiro and São Paulo), since the capitals have the same name as the state, our counting cannot tell them apart. Usually the state will have more occurrences than its capital, but in the case of Pernambuco, the opposite happens – its capital, Recife, has more occurrences.

Still, the generic frequencies of human-constructed clusters mirror some of the traditional observations of social scientists. For example Rio de Janeiro and São Paulo account for most of the locations mentioned in the DHBB entries, with the cluster (*rio_de_janeiro 8,154 rio 1,473*) with 9,627 occurrences being higher that the one for (*São Paulo 8,748*), possibly due to the fact that the entries cover the period when Rio was the capital of Brazil. Even the cluster for the word *Brasil* has fewer occurrences, 8,144. The cluster for the United States mentions (*estados_unidos 1,985 eua 785 nova_iorque 532 washington 823*) has 4,125 total occurrences, it is barely beaten by the occurrences of the cluster corresponding to the modern capital of Brazil since 1960, Brasília (*distrito_federal 2,784 brasília 1,799*), with 4,583 occurrences. This mirrors the importance of the United States in Brazilian politics. After the two main states/cities we have clusters corresponding to other states, e.g. (*minas_gerais 2,423 minas 553 belo_horizonte 1178*) and (*rio_grande_do_sul 2,500 porto_alegre 1,320*), and after that we have the whole Nordeste/Northeast as a single entity. We can also remark that the only other South-American country appearing at least 500 times (the Argentina/Buenos Aires cluster) has almost as many occurrences as the consolidated Nordeste. Given the massive immigration from Italy and Germany to the south of Brazil in the beginning of the 20th century, it is not surprising to see the number of times that these countries appear in the list, while the cluster for France and Paris primarily reflects the time when the elites of Brazil would study in Europe, specifically in Paris.

A strong feature of the political elite in Brazil, according to the social scientist Conniff [10], relates to their life experience abroad. Indeed, many of its members established contacts outside Brazil with many living for a while in other countries, either to study or to work, and in the times of the dictatorship, to seek political asylum. These are merely anedoctal observations, considering only the very top of the distribution of locations used. We need to engage in further discussions with social scientists to determine what kind of information would be useful to them and how to obtain it.

People When it comes to identifying named entities corresponding to people, we have some of the same problems which we had with locations. Some can be recognized in shortened forms ('Lula', 'Sarney', both former presidents of Brazil), others cannot ('Silva' has 945 occurrences, but it is unlikely that they correspond to the same person, as 'Silva' is comparable to the surname 'Smith' in England in terms of its popularity). Out of the 46 most frequent entities that are presumably people, we have some clear mistakes: some are places, actually names of Brazilian states, which could be mistaken for names of people (*Minas*, *Pernambuco*), some are political positions (*senador/senator*, *deputado/house representative*), while others are fixed multi-word expressions common in politics, such as *movimentação_financeira/financial moving* or *região_militar/military_zone*. These mistakes can be corrected, hopefully without much trouble. More worrying are the partial mappings, as for the name *joão_alberto*, which needs a surname associated to it to be identifiable. Finally the fact that some minor figures of Brazilian History such as *café_filho* or *góis_monteiro* appear in the list with such prominence is likely to be a result of processing mistakes. Of course, their prominence could also reflect a stronger bias of the DHBB to Brazilian history around the episode of the Revolution of 1930, but since this would be surprising, further investigation of the accuracy of the off-the-shelf FreeLing processing is required.

Organizations Detecting and classifying organizations shares some of the problems of detecting and classifying locations and people, in particular the need to cluster different expressions referring to the same entity, the need to deal with expressions that are not complete and the need to decide whether the expressions detected are reasonable or not. But the work on organizations brings a few other problems of its own. Acronyms are one of these. Acronyms can be very context-dependent. Out of 102 entities classified as "organizations", at least 20 are misclassified. Some are well-known state abbreviations (SP, MG, RJ etc.), some are names of ministries such as *Fazenda /Treasury* or *Interior/Homeland* mentioned only as such rather than with the proper full name (Ministério do Interior). Some are acronyms that are just hard to recognize for a casual reader. For example we detected an organization *const*: whenever a Constituent Assembly is convened with the purpose of reforming or drafting a new Constitution (in Brazil there were seven, so far), a fairly significant number of politicians - congressmen and senators - also accumulates the role of constituent, which is added to their biographies. The DHBB records about 1,100 constituents. In such contexts, the word should clearly not be considered an organization. Should this be considered a special named entity, particular to this domain? Perhaps, it is for the social scientists to decide, in general, which special named entities would be useful for them. Finally, some are indeed named entities, but not exactly organizations (e.g. *Estado Novo*, which is a name for a period in Brazilian History) and these need some classification in types too.

Dates Contrary to expectations, the most frequent dates in our processing of the DHBB did not provide us with sufficient information on periods of political turmoil in Brazilian history. The year that appeared most often (1930) is somewhat predictable, given that it is the year that Getúlio Vargas came into power through the *Revolução de 1930*. And the second most frequent year (1964) is also expected, given that it marks the year when the military came into power through their coup d'état, the *Revolução de 1964*, leading into the Brazilian military dictatorship. But the other most frequent years do not seem to mean much. The suicide of Vargas does not seem to be marked, nor is the huge public movement in the late 80's claiming for direct elections, called *Diretas Já*, which helped to oust the military. It is true that many of the observed years are associated with elections, which would be coherent with the nature of DHBB, but many of the important moments of recent Brazilian history, like the *AI5* (The Instutional Act 5) which marked the end of civil liberties in the country), the impeachment of President Collor and the associated scandal (which leads to at least one person *Paulo Cesar Farias* in the list of most frequent names) are not showing in our processing.

Given that the third most frequent year in our data from the DHBB is 1995, we asked a political scientist what could have been the cause. His explanation was as follows:

> This is the year when parliamentarians took office after the election of 1994. Also, 1995 was marked by a significant party migration movement among the politicians [11]. Many substitutes (suplents) took office too. All this may corroborate the position of the year 1995 in the ranking, since this information is propagated in the header of biographies and in texts as well. Of course, such phenomena also occurred in subsequent election years, but one must remember that the most significant upgrade of DHBB was in 2001 when all these congressmen and senators were included. Thus, although 1995 was not a year like 1930 and 1964, in political terms it was quite eventful.

On the one hand, this shows the need for data processing and, in general, digital organization experts to be in constant contact with the main intended users, the social scientists, as it seems that the processing of the headers has been given undue importance, if it makes it seem that 1995 was the third more important year in Brazilian History. On the other hand, what constitutes important events and years in history is definitely the preserve of the social scientists, but some correlation with frequency is expected. More investigation seems required. One option might be to move towards weighted frequencies, given that not every occurrence is equally meaningful.

Others Finally, we have also looked briefly at the so called *other* named entities, as provided by Freeling. These would, in theory, give us a glimpse of concepts in the DHBB that were important in Brazilian history, but were not people or organizations. As an example, *ato_institucional* is floated to the top of the frequency list of these other named entities (an *institutional act* was the kind of legal instrument used by the Executive branch during the dictatorship.) But for these other named entities we clearly cannot rely on out-of-the-box Freeling, but need instead to use human judgement to decide when the concepts obtained by the automatic processing make sense or not. It is also debatable which of these should be considered named entities: *lei/law* probably should not, while the *plano_nacional_de_desenvolvimento/National Plan for Development*, a specific planning tool of the government with several instances, probably should be considered a named entity.

V. Lexical-Semantic Analysis

Our long-term goal is to analyse the contents of the DHBB from a semantic perspective by identifying the specific meanings of words and capturing relationships and other *predica-*

tions. At this point, we only have taken very preliminary steps towards this goal, analysing nouns and verbs in the DHBB.

A. Nouns

We started with a simple frequency analysis of nouns, from which we can already discern the political and historical context. Many frequent nouns related to the government (*governo/government*, *estado/state*, *país/country*) and political positions (*presidente/president*, *deputado/congressman*, as well as activities (*mandato/mandate, partido/party, cargo/position, eleicao/election*) are among the most mentioned ones.

Looking at the most frequent nouns, we can also see that our previous work on NomLex-PT appears very useful here: whether a sentence mentions that someone *ganhou a eleição/won the election, fez estudos/did his studies* or someone *foi eleito/was elected, estudou/studied* should not make much of a difference. Semantically we are talking about the same event of winning an election/working towards a degree, whether we use the word *eleição/election, estudo/studies* or the words *eleito/elected, estudou/studied*. To be able to measure the relative importance of, say, elections/studies as a whole in the discourse of the DHBB, we need to be able to link the verb *eleger/elect* to the noun *eleição/election* (or *estudar/study* to *fazer estudos/do studies*), which is our reason for creating NomLex-PT. (The second example shows also that we have to face some of the difficulties associated with *light verbs* which are predominant in romance languages like Portuguese.)

Freely available lexicons of nominalizations exist for other languages, but we have not been able to find one for Portuguese, so we decided to create one such. We started from a simple translation from the English NomLex, added a translation of the French Nomage [12] and then used FrameNet [13] and the electronic resource Wiktionary to grow our stock of nominalizations. Then to ascertain the relative coverage of NomLex, we checked the most common nouns in the DHBB (more than 750 occurrences), removed the ones that were clearly not nominalizations (e.g. *ano/year, país/country, mês/month*) and ensured that all the nominalizations found were in NomLex-PT. We observed an error rate for NomLex of 25% and a small margin of processing errors, discussed in [9].

This small but clean and well-written corpus of historical biographies was the perfect setting for discovering highly used nominalizations in Portuguese (the DHBB entries are supposed to be easily read by undergraduate students) that were missing from our lexicon.

B. Verbs

In a frequency analysis of the verbs in DHBB, the most frequent ones, as expected, turn out to be light verbs and auxiliary verbs. But after these common ones (e.g.*ir/go*, *ser/be*, *ter/have*), it is possible to note a strong connection between the DHBB and the political universe. There is a consistent political bias, since politics is the main staple of the DHBB. Thus it is interesting to see that many of the most used verbs are related to political achievements (*assumir/take a position*, *eleger/elect*) or the evolution of a political career (*participar/participate*, *exercer/hold a position*, *integrar/be part of*). This can be observed in the great part of verbs with a substantial presence among DHBB entries.

However, given that we are mostly dealing with biographies of public figures, we have also a great proportion of verbs and nouns about life events like ‘being born’, ‘to wed’, ‘to graduate’, ‘to die’, ‘year’, ‘son’, etc. (*nascer, casar, formar, morrer, ano, filho*). As all of the entries are supposed to strictly follow the rules established for the structure of writing biographies described, where information regarding birth, education background, political experience, marriage and death are required, these occurrences seem predictable. From the data scientists’ perspective, this is one of the reasons why the DHBB corpus is interesting: there is some constraining of meaning, but it is humanly constructed. This is to be contrasted with *controlled languages* [14], where the syntactic and semantic forms are machine constrained. Those feel unnatural and unexpressive, while the DHBB text can be considered language ‘in the wild’, or as really written by humans.

C. OpenWordNet-PT and DHBB

OpenWordNet-PT provides us with a lexical ontology. Since OpenWordNet-PT is integrated with FreeLing, we are able to perform automatic word sense disambiguation to determine the most frequent disambiguated concepts in the DHBB. As a side effect, we obtain mappings from the original Portuguese tokens in the text to concept identifiers that can also be consulted in the English WordNet or in entirely other languages using UWN [7]. For instance, FreeLing correctly lemmatizes ‘nasceu’ to ‘nascer’ and then suggests WordNet 3.0 identifier 00360932-v, for which words in several languages as well as mappings to logical ontologies like SUMO [15] are available. In the long run, such mappings could enable us to perform logical inference, perhaps with a probabilistic logic, about events and situations reported in text.

In the short term, the DHBB corpus is already helping us to debug and improve OpenWordNet-PT, as we can use it as a reliable, albeit domain specific, corpus of correct and middle register usage of Portuguese. Since OpenWordNet-PT is automatically created from machine-based resources, we cannot fully guarantee either its accuracy or coverage. Checking prominent classes of words in the DHBB that are either not present or misrepresented in OpenWordNet-PT was one of the authors’ main goals when we started this project.

Our next step will be to go beyond disambiguating verbal concepts by additionally identifying their arguments. English resources like VerbNet [16] and FrameNet [13] are already connected to WordNet and thus OpenWordNet-PT. While equivalent Portuguese resources are scarce, we hope to be able to re-use some of the English resources. Additionally, we can exploit the fact that in the DHBB corpus, given its strict and enforced human guidelines on the writing, the bulk of the meanings would be concentrated in a not-so-huge collection of predications. This should make it much easier to canonicalize many of the most frequent predications, hopefully in flexible ways, so that we can help our 'users', the social scientists, to answer their questions.

D. Statistics

Table III describes the numbers of the full output of our light processing of the DHBB. Note that the full tokens include punctuation only tokens, data shown to indicate the size of the universe of discourse. Full analysis of these results is still under investigation.

Table III
DHBB PRELIMINAR STATISTICS

	total
entries of the DHBB	7517
sentences	307187
people	256588
places	132509
organizations	355388
other NEs	41813
nouns	1589494
verbs	892071
years	102332
full tokens	8881222

E. Case Study

Given the slightly surprising results of our processing of the dictionary entries, we thought it would be sensible to organize a small experiment to check the adequacy of the FreeLing tools to our special kind of historical corpus.

To this end we processed a small sample and manually verified what was missing or marked erroneously in this small sample. This is not a formal evaluation, simply a first checking of our tools and possibilities out-of-the-box.

It is possible to consider the tokens not from a global macroscopic perspective, using the whole DHBB dataset, but instead check whether they are being correctly classified within a single entry. The example below illustrates the idea. Tokens in green represent people, in blue organizations, in red verbs, in magenta locations, in yellow nouns, in cyan dates, and in black the other possibilities. We underlined the cases where we believed FreeLing misinterpreted the data.

"José Machado Coelho de Castro nasceu em Lorena (SP). Estudou no Ginásio Diocesano de São Paulo e bacharelou-se em 1910 pela Faculdade de Ciências Jurídicas e Sociais. Dedicando-se à advocacia, foi promotor público em Cunha (SP) e depois delegado de polícia no Rio de Janeiro, então Distrito Federal. Iniciou sua vida política como deputado federal pelo Distrito Federal, exercendo o mandato de 1927 a 1929. Reeleito para a legislatura iniciada em maio de 1930, ocupava sua cadeira na Câmara quando, em 3 de outubro, foi deflagrado o movimento revolucionário liderado por Getúlio Vargas. Ligado ao governo federal, encontrava-se ao lado do presidente Washington Luís, no palácio Guanabara, no momento de sua deposição no dia 24 de outubro."

It is interesting to notice some issues in FreeLing's output when analysing Portuguese texts. In this small snippet, it is possible to observe proper names, dates, and most of verbs being correctly classified. On the other hand, issues concerning the differentiation between locations and organizations are noticeable, as commented previously (e.g. *SP, Distrito Federal*). Also we can notice that in the future we should be recognizing so called multiword expressions, which is the case for *promotor público/public prosecuter* and *delegado de polícia/chief of police*. Some other issues are observed for adverbs (e.g. *ao lado/aside; no momento/at the moment*). In particular, proper handling of temporal propositional phrases is one of our next goals.

VI. Further Work

Our main goal with this project is to help historians have the possibility of asking specific questions about the semantics of the data in the DHBB. It is clear that the processing afforded by FreeLing plus NomLex-PT and OpenWordNet-PT is a first step, but we would like to be able to actually have the facts reported represented into some semantic representation, with the system being able to perform some form of logical inference over the data. We are not close to this yet.

As mentioned earlier, we are considering forms of semantic role labelling as a next step in this project. It is clear that discovering the relationships described in the entries (who did what to whom and when and why) is crucial for this project. Given our domain interest in history, it seems that particular attention should be given to deciding on the marking of temporal connections between sentences, which may involve devising and implementing some theory of causality. But for all the interesting work that lies ahead, we would like to take our cues from the historians who have generously discussed their questions and problems with us.

We wish to see the DHBB as a source for further research and analysis about the history and politics of contemporary Brazil. When historians and data scientists gather together in cooperation, it is the former who – besides supplying the corpus and the data — have the task of addressing which questions and issues they would like to see solved by looking at those data. We must discover, then, the existing potential in these resources and develop intuitions of where to go with the help of data extraction, clustering, natural language processing, and other techniques.

At the end of the 1980s, a study conducted by Michael Conniff and Sonny Davis [17] with a sample of 7% of the entries (about 250 entries, then) allowed them to locate important changes on social classes, on regional origins and on economic resources of the national political elite. With the help of human assistants, they made a selection of individuals who occupied positions in the Executive branch, read the texts and manually extracted the desired information using SPSS[2] to consolidate the data. The results, extremely interesting, brought considerations like the one below.

> Considering the political elite of individuals born between 1900 and 1950 who occupied positions in

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

> the Executive Branch, it is possible to observe the trend and behavior of the variable "education" as follows:
>
> 'The majority university diplomas obtained by members of the elite were on law (44%), one quarter of which were located outside the original place of birth. The second most recurrent was reached in military educational establishments (32%). The diplomas in engineering and medicine stood at third and fourth place (12% and 5%, respectively). Over time the most significant change was observed on the decline in military training for those born after 1920, from 37% to 10%.' [10]

There are many questions that researchers have in mind when looking at the informational richness inherent to the DHBB. A small demonstration can be seen through the following only semi-humorous question we asked some historians and political scientists: "Imagine that you earned a robot assistant capable of quickly reading all the entries of the DHBB and undertaking any kind of data prospecting you wish for. If this were possible, what questions would you like to see answered?" The excitement of having something desired but still far from reality brought forth thousands of interesting questions. Questions that, without digital aids, the researchers would only have a chance of answering after thoroughly reading the whole body of texts. Here are some of their questions:

- Who are the politicians who died from non-natural causes or committed suicide?
- Did women who live in Rio de Janeiro and occupied high positions in the Executive branch between the decades of 1960 and 1980 attend the same academic circles or the same intellectual environment?
- How the military presence in the political elite was brought about? By bureaucratic or revolutionary means?
- What are the most frequent paths among politicians who occupied the president and governor's chairs? How many passed through the Legislative Branch? Does it change much from one period to another?
- What was the educational profile of the ministers of the Supreme Military Court between 1934 and 2010?
- Can you group data about education in accordance with the political and historical periods of the Republic? I.e. how did the academic training of policy designers evolve during the Old Republic (1889-1930), the 1st Vargas Period (1930-1945), the Democratic Governments (1945-1964), the Military Regime (1964-1985) and the New Republic (1985-now)?

Driven by these questions and other challenges, our proposal is to keep working on initiatives to build bridges between the communities of data scientists and social scientists, that will make it possible to retrieve more information, more accurately, in more useful and actionable ways.

References

[1] A. A. de Abreu, F. Lattman-Weltman, and C. J. de Paula, Eds., *Dicionário Histórico-Biográfico Brasileiro pos-1930*, 3rd ed. Rio de Janeiro: CPDOC/FGV, 2010, available at http://cpdoc.fgv.br/acervo/dhbb.

[2] C. Fellbaum, *WordNet: An electronic lexical database*. The MIT press, 1998.

[3] R. R. Korfhage, *Information Storage and Retrieval*. Wiley, 1997.

[4] A. Singhal, "Modern information retrieval: A brief overview," *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, vol. 24, no. 4, pp. 35–43, 2001.

[5] L. Padró and E. Stanilovsky, "Freeling 3.0: Towards wider multilinguality," in *Proc. of the 8th Intern. Conf. on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012, pp. 23–25.

[6] V. de Paiva, A. Rademaker, and G. de Melo, "OpenWordNet-PT: An open Brazilian wordnet for reasoning," 2012.

[7] G. de Melo and G. Weikum, "Towards a universal wordnet by learning from combined evidence," in *Proc. of CIKM 2009*. New York, USA: ACM, 2009, pp. 513–522.

[8] F. Bond and K. Paik, "A survey of wordnets and their licenses," in *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue, 2012, 64–71.

[9] V. D. Paiva, L. Real, A. Rademaker, and G. D. Melo, "Nomlex-pt: A lexicon of portuguese nominalizations," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, N. C. C. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds. Reykjavik, Iceland: European Language Resources Association (ELRA), may 2014.

[10] M. Conniff, "O DHBB e os brasilianistas," in *CPDOC 30 Anos*, E. FGV, Ed. Rio de Janeiro: Editora FGV/CPDOC, 2003.

[11] C. R. F. d. Melo, "Partidos e migração partidária na câmara dos deputados," *Dados*, vol. 43, 2000. [Online]. Available: \url{http://goo.gl/Imc4Qy}

[12] A. Balvet, L. Barque, M. H. Condette, P. Haas, R. Huyghe, R. Marn, and A. Merlo, "Nomage: an electronic lexicon of french deverbal nouns based on a semantically annotated corpus," in *Proceedings of the First International Workshop on Lexical Resources*, Ljubljana, Slovenia, 2011.

[13] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The Berkeley FrameNet project," in *Proc. COLING-ACL 1998*, 1998, pp. 86–90.

[14] S. O'Brien, "Controlling controlled english. an analysis of several controlled language rule sets," *Proceedings of EAMT-CLAW*, vol. 3, pp. 105–114, 2003.

[15] I. Niles and A. Pease, "Towards a standard upper ontology," in *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, C. Welty and B. Smith, Eds., Ogunquit, Maine, October 2001, pp. 17–19, see also http://www.ontologyportal.org.

[16] K. Kipper, A. Korhonen, N. Ryant, and M. Palmer, "Extending verbnet with novel verb classes," in *Proceedings of LREC*, vol. 2006, no. 2.2. Citeseer, 2006, p. 1.

[17] F. D. McCann and M. L. Conniff, *Modern Brazil : elites and masses in historical perspective / edited by Michael L. Conniff and Frank D. McCann*. University of Nebraska Press Lincoln, 1989.