Text Mining for History: first steps on building a large dataset

Suemi Higuchi, Cláudia Freitas, Bruno Cuconato, Alexandre Rademaker
FGV/CPDOC and PUC-Rio, PUC-Rio, FGV/EMAp, IBM Research and FGV/EMAp
suemi.higuchi@fgv.br, claudiafreitas@puc-rio.br, bcclaro@gmail.com, alexrad@br.ibm.com

Abstract

This paper presents the initial efforts towards the creation of a new corpus on the history domain. Motivated by the historians' need to interrogate a vast material in a non-linear way, our approach privileges deep linguistic analysis on an encyclopedic-style data. In this context, the work presented here focuses on the preparation of the corpus, which is prior to the mining activity: the morphosyntactic annotation and the definition of semantic types for entities and relations relevant to the History domain. Taking advantage of the semantic nature of appositive constructions, we manually analyzed a sample of eleven hundred sentences in order to verify its potential as additional semantic clues to be considered. The results show that we are on the right track.

Keywords: digital humanities, text mining, corpus annotation, appositives

1. Introduction

Language is a rich repository of information about our practices, constituting raw material for research in Human and Social Sciences. In close connection with Computational Linguistics, Humanities and Social Sciences, the growing field of the Digital Humanities has at its disposal tools and resources that offer new ways to explore many available corpora.

In this paper we present our initial efforts towards the creation of a resource dedicated to text mining in the history domain. The mining strategy is linguistically motivated: inspired by (Hearst, 1992) we assume that certain semantic relations have a linguistic realization, and therefore the inclusion of linguistic metadata such as part-of-speech, lemma, and syntactic information in the corpus is essential. The target of the mining - the corpus - is the *Dicionário Histórico-Biográfico Brasileiro* (Brazilian Historical-Biographical Dictionary), DHBB for short, that contains almost 12 millions tokens in about three hundred thousand sentences.

The DHBB is a reference work, written by historians and social scientists and published by the Contemporary Brazilian History Research and Documentation Center (CPDOC) of Getulio Vargas Foundation (FGV). It contains almost eight thousand entries with information ranging from the life and career trajectories of individuals to the relationships between the characters and events that the country has hosted. The primary motivation to mine the DHBB came from the need to query the material looking for information that requires almost total reading of the whole body of texts. Examples of such inquiry could be the kinship (or personal) relationship between politicians and their connection to entities such as institutions, movements, events or places throughout their public life. That is, we aim to construct a resource able to answer questions such as "Which politicians were born before the 1960s, had military training and held a position in the Executive Branch?".

We are aware of the vast amount of knowledge spread around the entries in a non-linear way. After all, dictionaries and encyclopedias are made to be consulted and not to be read linearly. In this context, the focus of this paper is to report the first efforts related to the preparation of the material – in particular, the morphosyntactic linguistic annotation and the definition of semantic types for entities and relations relevant to the History domain, taking the appositives as important syntactic relation to observe when annotating semantic relations.

Our main purpose is not only to mine the DHBB, but to create a public corpus to foster Portuguese NLP in general, and NLP in the history domain, in particular. Most large Portuguese annotated corpora are composed of newspaper texts; the DHBB entries, on the other hand, are written in encyclopedic style, and this "novelty" can be a challenge for automatic parsers.

The paper is organized as follows: Section 2. presents the preparation of the corpus relating to the morphosyntactic annotations and the motivation for the comparison exercise performed between two parsers: UDPipe and PALAVRAS. In Section 3. we present the entities types and relations relevant to the History domain. In Section 4. we detail the manual analysis that we conducted of the appositive relations between entities, the evaluation of the outputs generated by the parsers and the revision of entities identification/segmentation of proper names in a sample of the corpus. Finally, concluding remarks are summarized in Section 5..

2. Corpus Preparation

The first edition of the DHBB dates from 1984 in printed version only, and it was in 2010 that its content was fully made available on the Internet[1]. Since its beginning, CPDOC has developed an internal information system to maintain the data through forms and reports that interact with a relational database. The database structure can be summarized as one main table that contained a text field with the entries encoded in HTML and some metadata: basically, the name of the entry and its nature (whether biographical or thematic). This structure showed to be quite limiting when it concerns maintenance and improvements on the dictionary. These issues are described in details in (Paiva et al., 2014) and were the reason for our proposal of maintaining the entries as text files using a lightweight human-readable markup syntax, like YAML (Ben-Kiki and Evans, 2005) and Markdown (Gruber, 2004). A consider-

[1] Available at `http://cpdoc.fgv.br/acervo/dhbb`

able effort was then made to bring up this structure. The decision to adopt plain text files was justified by clear reasons: easiness of maintenance using any text editor (tool independence); conformity to long-term standards by being software and platform independent; easiness to exploit the possibilities of DHBB's files as a resource for NLP; enrichment of the entries with metadata of any kind at any time, even those extracted from natural language processing. The data is freely available at `https://github.com/cpdoc`.

Among the many linguistic metadata that we are adding to DHBB corpus, one important annotation layer is the syntactic analysis.[2] The syntactic analysis is being done according to the Universal Dependencies standards (Nivre et al., 2016). The Universal Dependencies (UD) project,[3] in its ambitious and encompassing mission of affording a single set of tags and parallel analyses common to several different languages, provides a multilingual natural language processing (NLP) framework. The general philosophy is to provide a universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages, while allowing language-specific extensions when necessary. An example of such annotation is given in Figure 1, showing the main grammatical relations involving a verb, an oblique agent and an appositive.

In order to parse DHBB, in this first stage we run both PALAVRAS (Bick, 2000), a rule-based multi-level constraint grammar parser developed specifically for the Portuguese language, and UDPipe (Straka and Straková, 2016), a machine learning pipeline for tokenization, tagging, lemmatization and dependency parsing. UDPipe follows the UD's guidelines, being language-independent, and can be trained given annotated data in CoNLL-U format.[4]

The motivation for the double processing is twofold: first of all, we believe that comparing the outputs of different systems is a way to optimize the linguistic human revision, as suggested in (Truggo, 2016). Additionally, we aim to compare linguistic analysis of both systems in a genre (encyclopedic) unusual to these parsers.

On the whole, the DHHB corpus comprises: automatic morphosyntactic annotations given by the parsers for the whole corpus, manual entity relations annotations for the golden sample, and an entity lexicon built semi-automatically from lexical-syntactical patterns, taking advantage of the highly predictable written style of the DHBB.

3. Entities and Relations

Entity recognition is a crucial task for text mining since its main focus is on instances of general semantic types like person, location, time and organization. Our definition of entity closely follows the ACE (Automatic Content Extraction) proposals (Doddington et al., 2004), capturing all kinds of information that can identify something or someone relevant, whether it's a proper name or not. In an entry about *Revolução de 1930* (Revolution of 1930), for instance, we intend to recover data about this specific event even when it is referred as *revolução* (revolution) as in "Essa carta pode ajudar no esclarecimento de um ponto importante das articulações da *revolução*, pois a bibliografia sobre o período refere-se a dois encontros entre Vargas e Prestes" (This letter can help clarify an important point of the articulations of the *revolution*, since the bibliography on the period refers to two meetings between Vargas and Prestes).[5]

To elicit the semantic types relevant for the history domain, we combined knowledge from domain experts and a corpus driven approach based on a wide reading of entries, aimed at validating and increasing the initial proposed classes. As a result, we conceived seven classes, presented in Table 1.

Classes	Examples
`PER` (person)	*Getúlio Vargas, Lula, presidente*
`ORG` (organization)	*Petrobras, Partido Democrático Social, PDS*
`POL` (political formulation)	*Plano Collor, Programa de Estabilização Monetária, AI-5*
`EVN` (event)	*Revolução de 1930, Atentado do Riocentro*
`LOC` (local)	*São Paulo, palácio Guanabara*
`DOC` (document)	*Diário pessoal de Getúlio Vargas*
`TME` (time)	*Janeiro de 2001, 1927 a 1929*

Table 1: Entity classes for DHBB

Inspired by the set of relations proposed by the Second HAREM task (Freitas et al., 2008) we devised our own set of relations to connect the entities. During the process of text analysis, in particular looking at appositives occurrences, a few other relations were identified as relevant to our goals. Table 2 presents the final list of relations and examples.

4. DHBB: Hands-On

For the work presented here, whose primary purpose is to offer some insights to the text mining in the history domain, we selected a sample of 35 dictionary entries containing 38,554 tokens in 1,115 sentences. To convert our sample in a golden set, we conducted a manual revision of the appositive relations between entities: (i) revising the segmentation of the entities names; and (ii) manually identifying the induced semantic relationship between the entities.

In the following, we detail each of these steps. Along the process, we also analyzed (i) the quality of the automatic parsing as to appositive structures; and (ii) the impact of named entities domain lexicon in proper names segmentation. Table 3 presents some features of the sample.

[2]The DHBB files with linguistic metadata are available in `https://github.com/cpdoc/dhbb-nlp`.

[3]http://universaldependencies.org

[4]http://universaldependencies.org/format.html

[5]We are yet to address the co-reference resolutions.

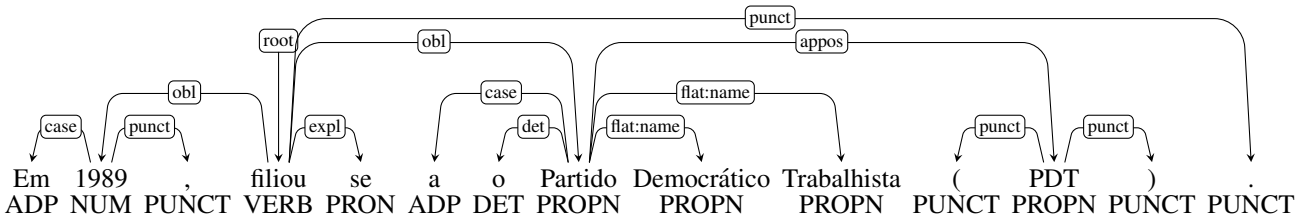


Figure 1: *In 1989, [he] affiliated with the Democratic Labour Party (PDT).*

Relations	E1	E2
`ident` (coreference)	*Partido dos Trabalhadores*	*PT*
`role`	*Alberto Coelho*	*president*
`loc` (local)	*port of Alcantara*	*in Lisbon*
`part`	*Porto Seguro*	*BA*
`date`	*promulgation of Nova Carta*	*18/9/1946*
`link-inst` (institutional relation)	*Vandilson Costa*	from Partido Comunista do Brasil
`link-fam` (family relation)	*Nilo Augusto*	*son of Gercino Coelho and Eunice Coelho.*
`link-pers` (personal relation)	*Orígenes Lessa*	*friend of his brother Fúlvio*
`attrib` (attribute)	*João Abdalla and Amélia Abdalla*	*of Arab origin*
`participant`	*Getulio Vargas*	*in the Revolution of 1930*
`context` (happens)	*XXXVIII ministerial meeting*	*of General Agreement on Tariff and Trade*

Table 2: Relations between entities

Freq	information
38,554	tokens
1,115	sentences
472	sentences with at least one appositive
796	appositives
10	types of semantic relations

Table 3: Details of the revised sample of 35 DHBB entries

Finally, we should note that we have not revised all syntactic annotations in the sample. We have focused our attention only on the names segmentation and the appositives relations. It is an ongoing work the release of a completely revised syntactic analysis of the corpus.

4.1. Appositives

Appositives are syntactic relations especially productive for text mining, with contributions to the building of semantic lexicons, noun phrase co-reference resolution and information extraction from texts (Freitas et al., 2006). They provide descriptive information about the head noun, thus enriching its characterization: when a given noun is tagged as an appositive, a relationship with another term is derived. Appositive relations induce many different semantic relations between entities. In Table 2, the examples of relations `ident`, `role` and `link-fam` all appear on the text as appositives relations between the entities.

Using the output of UDPipe, the revision process was steered in two steps. First, we revised the entities segmentation/identification. Then, we used the PALAVRAS output to check for any missed or incorrect appositive annotation from UDPipe. Along these steps, we annotated the semantic relations between entities expressed in appositive constructions within our golden sample.

The explicit semantic nature of appositives led us to a semantic strategy for revision the parser analysis. That is, we extracted from the parsed sentences the triples formed by the `appos` relation – the linearization of the noun phrases that have their heads connected by an `appos` relation. The extracted triples can be trivially analyzed by humans and if abnormal noun phrases appear it indicates a possible parser mistake.

From table 3, we know that 796 appositive relations were found in 472 sentences. Considering that the UD schema can provide up to 35 possible syntactic relations, the frequency of appositives compared to the other relations can be a clue to depict other linguistic analyses within the corpus. For instance, we observe that core arguments such as (`obl`, `obj`) are far more frequent than subjects (`nsubj`), the reason lies on the style of the narrative that privileges the use of *implicit subjects*, a construction that does not exist in the English language but that is very common in Portuguese written texts (e.g, Figure 1). Table 4 shows the distribution of the first fifteen syntactic relations in the golden sample.

Ten different types of semantic relations from our tagset were identified among the 796 appositive occurrences. Table 5 presents the distribution of semantic relations assigned by the human reviewer for each appositive relation on the golden sample.

Not surprisingly, the most common types of semantic relationship that the appositive constructions reveal are those of *role* and *ident*. On the other hand, personal relations such as friendship or fellowship are almost never made explicit

Freq	information
7,710	case
5,599	det
4,755	punct
3,967	nmod
2,693	obl
1,905	flat:name
1,418	amod
1,115	root
1,107	obj
1,013	conj
899	nsubj
804	cc
796	**appos**
752	advmod
544	compound

Table 4: Distribution of the 15 most frequent syntactic relations occurring in the sample

Num	semantic relation	%
300	role	37.7
200	ident	25.1
73	attrib	9.2
73	date	9.2
65	link-fam	8.2
62	part	7.8
11	link-inst	1.4
6	loc	0.8
5	other	0.6
1	link-pers	0.1

Table 5: Frequency distribution of types of relations in the revised sample

in DHBB, at least not through appositive constructions.

4.2. Evaluation of systems performance

The PALAVRAS system recognized 797 cases of appositives and UDPipe 954 cases.[6] After manual revision, our sample contains 796 occurrences. Although these numbers may suggest that PALAVRAS achieved a better score than UDPipe, these numbers taken globally do not reveal the effective quantity of mistakes that were corrected. Below we elaborate on the comparison of the golden set (the revised sample) with the UDPipe output. Unfortunately, since PALAVRAS analysis follows an entirely different tagset and directives, we are not able to make a detailed comparison of both systems. However, during the revision underlying the construction of the golden sample, we observed that PALAVRAS also produced many incorrect analyses.

When comparing UDPipe's output with the revised sample, we distinguished the following cases:

AllCorrect correct identification of the arguments of the relation and correct identification of appositive. See Figure 1.

ErrDepRel correct identification of the arguments of the relation but incorrect identification of appositive. Figure 2a.[7]

ErrHead incorrect identification of the arguments of the relation but correct identification of appositive, Figure 2b.

FullErr incorrect identification of argument and relation, Figure 2c.

MissingAppos an appositive relation was not detected, Figure 2d.

Table 6 presents the results of the qualitative analysis of UDPipe performance on appositive structures.

Num	Errors/success	%
492	AllCorrect	53.1
9	ErrDepRel	1
175	ErrHead	18.9
203	ErrNotAppos	21.9
47	ErrMissingAppos	5.1

Table 6: Frequency distribution of UDPipe's errors concerning appositive relations

Appositives are tricky linguistic structures to be parsed automatically, since its main formal clue, punctuation, can be easily confused with coordination. For example, in the sentence

> Entre 1959 e 1960, coordenou o setor financeiro da campanha eleitoral do marechal Henrique Teixeira Lott, candidato à presidência da República apoiado pelo PSD e o PTB. (*Between 1959 and 1960, he coordinated the financial sector of the election campaign of Marshal Henrique Teixeira Lott, candidate for the presidency of the Republic supported by the PSD and PTB*)

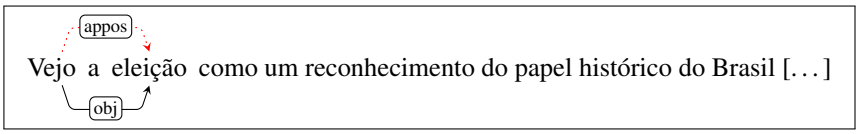
UDPipe (erroneously) analyzed "candidato" in coordination with "setor". Also, in the sentence

> ...votou a favor da emenda constitucional que previa a reeleição de presidente da República, governadores e prefeitos, ...(*voted in favor of the constitutional amendment that foresees the reelection of the president of the Republic, governors and mayors,*)

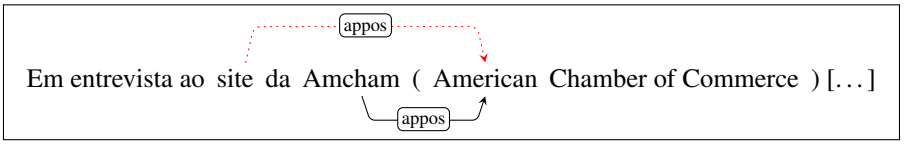
both PALAVRAS and UDPipe were mistaken in identifying an appositive structure between "governadores" and "reeleição" when it is a clear case of coordination.

[6] PALAVRAS uses two tags to indicate the general idea of appositives, we have considered both tags.

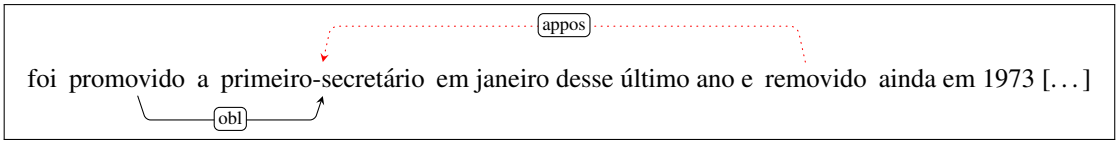
[7] For the rest of this paper, edges above a sentence in red dotted lines represent incorrect analyses, while edges below represent the correct analysis.



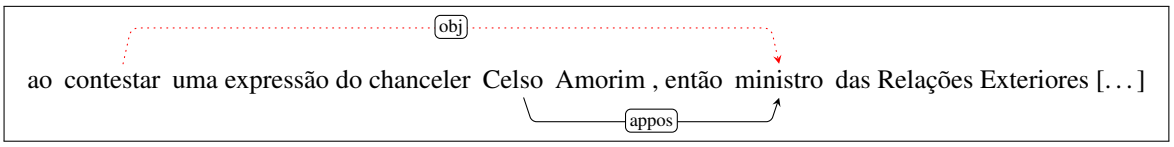
(a) *[I] see the election as a recognition of Brazil's historic role*



(b) *In an interview to Amcham's (American Chamber of Commerce) website*



(c) *[he] was promoted to first secretary in January of the previous year, and removed in 1973*



(d) *when challenging chancellor Celso Amorim's expression, then Foreign Affairs minister*

4.3. Proper nouns identification and segmentation evaluation

Prior to semantic classification of the named entities (NE) we need to correctly identify them. By the highly idiosyncratic nature of proper nouns, errors resulting from wrong segmentation are usual. As an example, PALAVRAS and UDPipe considered *Ministério das Minas e Energia* (Ministry of Mines and Energy) as two separated names: *Ministério das Minas* and *Energia*. Regarding person names, UDPipe split the last name of the person *José Afonso de Melo*, as a noun modifier of the first, and not as part of the whole name. In this case, PALAVRAS did it right, joining the tokens in a single unit.

Each parser has its own way of performing the proper noun segmentation and we tried to reduce the errors by creating domain lexicons from external resources. For the lexicon of *person names*, for instance, we have used both DHBB metadata and a list provided by the personal archives system from CPDOC (Rademaker et al., 2015), being possible to gather 18,171 names. As to the *organizations*, we have almost entirely used pattern recognition in the corpus to extract the names: with AntConc (Anthony, 2016) we searched for simple patterns like *presidir o [A-Z]* (to chair the [capital letter]) or *estudar em [A-Z]* (to study at [capital letter]). This process lead to a lexicon of 3,637 entities.

In the entire corpus we found 83,898 person names (7,514 of them unique) that exist in the lexicon being mentioned on the text, which represents 42% of the whole list. Concerning the organizations we found 69,775 names (3,029 of them unique) occurring in the corpus, which represents 83% of the lexicon. The reason why we have a higher number of organization names matches is due the approach used to construct the list, as explained above, which make use of lexical patterns and concordance lines to extract the names from the corpus.

Concerning our golden subset we found 430 persons (219 of them distinct) persons mentioned on the text. As to the organization lexicon, we found 360 organizations (116 of them distinct) organizations occurring in the corpus.

We know that the use of lexicons has limitations such as limited coverage and variation in the writing of names, i.e. the same person can be mentioned in different ways ranging from the complete full name to the nickname. On the other hand, we believe that the incorporation of lexical entries, associated with semantic classes, are a simple and effective method to bootstrap the creation of lexical-syntactic patterns, crucial for semantic annotation between entities.

Some studies demonstrate positive results when adopting similar approach of using lexicons. In (Florian et al., 2003), the authors investigated the combination of a set of diverse statistical named entity classifiers applied to an English corpus: when no lexicons (gazetteers) or other additional training resources are used, the combined system attains a performance of 91.6 F1 on the English development data; but after integrating gazetteers containing some 50 thousand names of cities, 80 thousand proper names and 3,5 thousand organizations, the F-measure error was reduced by a factor of 15 to 21%.

We then evaluated the impact of using the lexicon for automatic post-processing the UDPipe before comparing it with the golden sample.

Entity name recognition was done in a greedy way. If both "José Machado Coelho de Castro" and "José Machado" were in the lexicon, and the former were in a sentence, only the former would be recognized. In fact, even if "Machado Coelho" or "Castro Abreu" were in the lexicon and the phrase were "José Machado Coelho de Castro Abreu nasceu em 1931", only "José Machado de Castro" would be recognized, provided "José Machado Coelho de Castro Abreu" is not in the lexicon.

Thus, we have the following: 790 mentions of proper

names from the lexicons were found in the golden set sentences, with the most frequent name occurring 53 times. Although we have made corrections in 460 tokens, only thirty of the affected tokens had been marked with an `appos` relation.

In 18 of these cases, the wrong segmentation of the name had caused an error in the syntactic dependence of the appositive token, and this has been fixed with the incorporation of the lexicon, see the example of Figure 3.

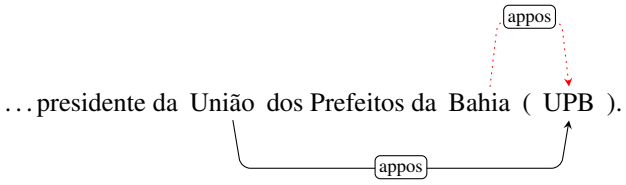


Figure 3: president of the Mayor's Union of Bahia State (UPB)

And in the remaining 12 cases, one of the name's token had been erroneously marked as having an appositive relation with the first token. This was also been fixed with the lexicon, suggesting a relevant role of the lexicon for the syntactic analysis as shown in the example of Figure 4.

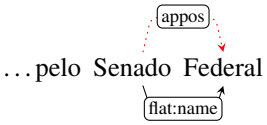


Figure 4: by the Federal Senate

5. Concluding Remarks

In this paper, we present the first efforts towards the creation of an annotated corpus for the history domain. Motivated by the historians' need to interrogate a vast text material, our approach privileges linguistic analysis, as opposed to techniques such as topic modeling, which we believe to be complementary.

In this context, a crucial step is to prepare the material that will be mined. In our case, the preparation includes the annotation of morphosyntax, entities and semantic relations. Although the morphosyntactic annotation is already being performed automatically, the results are not reliable, at least in relation to appositives, as we understand from our analyses in Table 6.

Although highly informative for text mining, appositives seem to be quite difficult structures for systems. On the other hand, it is worth remembering that UDPipe was trained on Bosque-UD (Rademaker et al., 2017), a corpus of a different genre (newspaper), not too big (227,842 tokens), and, ironically, built upon manual revision of PALAVRAS analyses. From a linguistic point of view, the apposition is a syntactic relation only apparently simple (and this point is signalled in the UD guidelines dedicated to *appos*), and, to a Brazilian grammarian, it is "an obscure object" (Perini, 1996).

As to proper names and the lexicons, to compile a comprehensive list of names, we faced difficulties that are particular to the corpus and to the Brazilian Portuguese. The first challenge is related to the DHBB guidelines and has to do with normalization of person names. 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 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 of that time: the letters 'Y' and 'W' are replaced by 'I' and 'V' ('Darcy' becomes 'Darci', 'Oswaldo' becomes 'Osvaldo'), in some cases 'Z' becomes 'S' (then 'Souza' becomes 'Sousa' and 'Menezes' becomes 'Meneses'). Such rules may appear unusual and dispensable in modern times when data is digitized and expected to be retrieved by search engines capable of answering more advanced requests with wildcard, range, and fuzzy queries. In later versions these normalization rules were dropped and therefore entity names across entries might be inconsistent. Similar to this issue is the Brazilian orthographic reform that took place in 2009. Some of the changes made include extinguishing the use of some hyphens and accents, like "infra-Estrutura" (*infrastructure*) that has become "infraestrutura" and "assembléia" (*assembly*), now "assembleia". All these variations must be in the lexicons in order to improve the parser processing and semantic classification.

In addition, there are some domain's particularities, like entity types such as "Policy Formulation", that would hardly be included in general-purpose NE systems. Another important issue that we have glimpsed but did not focus on in this study has to do with co-reference resolution. Cases like Figure 1 illustrate the non-explicitness of the subjects that is very common in the sentences of the DHBB. The syntactic structure analyses in conjunction with clues like the biographee's name given in the first sentence, can appear in the strategies to be adopted. There is a long way to go.

6. Bibliographical References

Anthony, L. (2016). Antconc version 4.4.1, computer software.

Ben-Kiki, O. and Evans, C. (2005). Yaml ain't markup language (yaml™) version 1.1.

Bick, E. (2000). *The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press.

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S., and Weischedel, R. M. (2004). The automatic content extraction (ace) program-tasks, data, and evaluation. In *LREC*, volume 2, pages 837–840.

Florian, R., Ittycheriah, A., Jing, H., and Zhang, T. (2003). Named entity recognition through classifier combination. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume*

4, CONLL '03, pages 168–171, Stroudsburg, PA, USA. Association for Computational Linguistics.

Freitas, C., Duarte, J. C., Santos, C. N., Milidiú, R. L., Rentería, R. P., and Quental, V. (2006). A machine learning approach to the identification of appositives. In *Advances in Artificial Intelligence-IBERAMIA-SBIA 2006*, pages 309–318. Springer.

Freitas, C., Santos, D., Gonçalo Oliveira, H., Carvalho, P., and Mota, C. (2008). Relações semânticas do rerelem: além das entidades no segundo harem. *Linguateca*, pages 77–96, Jan.

Gruber, J. (2004). Markdown language.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Paiva, V. D., Oliveira, D., Higuchi, S., Rademaker, A., and Melo, G. D. (2014). Exploratory information extraction from a historical dictionary. In *IEEE 10th International Conference on e-Science (e-Science)*, volume 2, pages 11–18. IEEE.

Perini, M. A. (1996). *Gramática descritiva do português*. Editora Ática.

Rademaker, A., Oliveira, D. A. B., de Paiva, V., Higuchi, S., e Sá, A. M., and Alvim, M. (2015). A linked open data architecture for the historical archives of the getulio vargas foundation. *International Journal on Digital Libraries*, 15(2-4):153–167.

Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., and de Paiva, V. (2017). Universal dependencies for portuguese. In *Proceedings of the International Conference on Dependency Linguistics*, Pisa, Italy.

Straka, M. and Straková, J. (2016). UDPipe. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

Truggo, L. F. (2016). Classes de palavras - da grécia antiga ao google: um estudo motivado pela conversão de tagsets. Master's thesis, Programa de Pós-Graduação em Estudos da Linguagem, Departamento de Letras PUC-Rio.