Distant reading Brazilian politics

Suemi Higuchi (FGV, PUC-Rio), Diana Santos (Linguateca, ILOS/UiO),
Cláudia Freitas (Linguateca, PUC-Rio), Alexandre Rademaker (IBM, FGV)

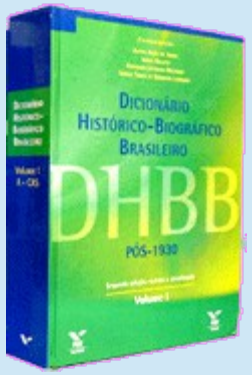
DHN 2019 Conference - Copenhagen

About the research

- Goal: to extract structured information from large volumes of text in Portuguese;
- Resource: *Dicionário Histórico-Biográfico Brasileiro* (DHBB);
- Biographical and thematic entries;
- Domain: Contemporary political history of Brazil (1930-);
- A reference work for historians, political scientists, journalists, etc.

1984

Print version

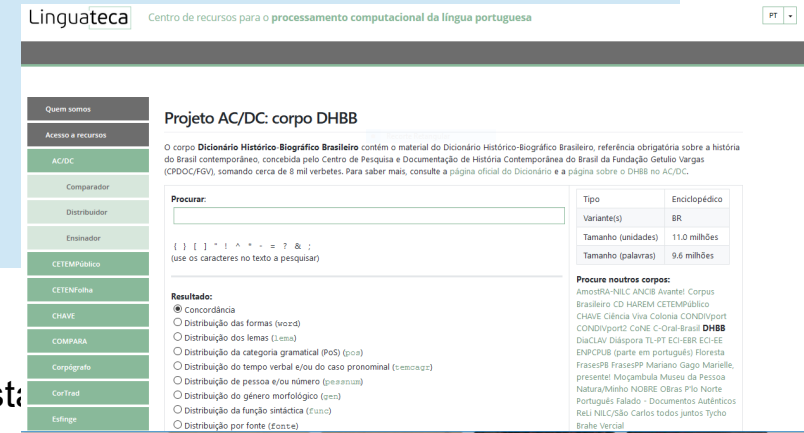


2001

Database format available on internet

2018

Corpus version available from Linguateca



An overview of DHBB

Biographies of	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 (members of t Deputies)	
Militares (Army officers)	
Participantes de revoluções (revolu	
Jornalistas (journalists)	

Entries	7,421
· biographical	6,660
· thematic	761
Tokens	10,079,830
Sentences	314,168

Two-fold motivation for our work

1. From the DHBB point of view
   Improve the usability of DHBB:

   "is it possible to extract some information without having to read the entries individually?"

   - *Among the political elite included in the DHBB, which of them died of non-natural death?*
   - *How many (and who) belong to a family of politicians?*
   - *What is the most common educational background of the members of the Legislative Branch?*

   Have a coherent view of the whole of Brazilian political landscape described in DHBB

2. Increase the available resources for the Portuguese language and let other corpora benefit from analysis done for DHBB

About the DHBB textual material

- Predictable and ‘well behaved’ writing
- Guidelines that contain samples of ready-to-use sentences, verb tenses to adopt, paragraph structure and descriptive sequences to follow

GOULART, João
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.
....

GOULART, João
João Belchior Marques Goulart was born in São Borja (RS), on March 1, 1919, son of Vicente Rodrigues Goulart and of Vicentina Marques Goulart. From the time he was a child he was given the nickname Jango, common in the south of the country.
....

In the following paragraphs: academic background, professional activities, political trajectory, date of death (if deceased), names of spouses and children.

Corpus analysis

- Conversion of DHBB into a corpus →plain text with some metadata, encoded in Open Corpus Workbench ;
- Linguistic annotation with the PALAVRAS parser and other modules, covering morphosyntax and some semantic annotation (e.g. about proper names)

Proper names	**types**
1,6 million tokens	ca. 48,500 persons
(118,000 different ones)	ca. 27,500 organizations
	ca. 5,000 places
	ca. 36,900 others

Entity grounding Making sense of proper names

- Correspondance between names that refer to the same person:

Lemma (nicknames, part of the name, mispellings...)	Entry name in DHBB (entity)
Bolsonaro	Jair Messias Bolsonaro
Getulio Vargas	Getúlio Dornelles Vargas
Jango	João Belchior Marques Goulart
Lula	Luis Inácio da Silva

- Iterative process:

Proper names (tokens)	Number of rules	Identified entry names (tokens)	Entity grounding	Corpus version
404.245	115	89.937	22%	1.0
404.245	187	147.085	36%	1.2
478.333	271	166.059	34%	2.0
478.333	312	239.106	50%	2.8

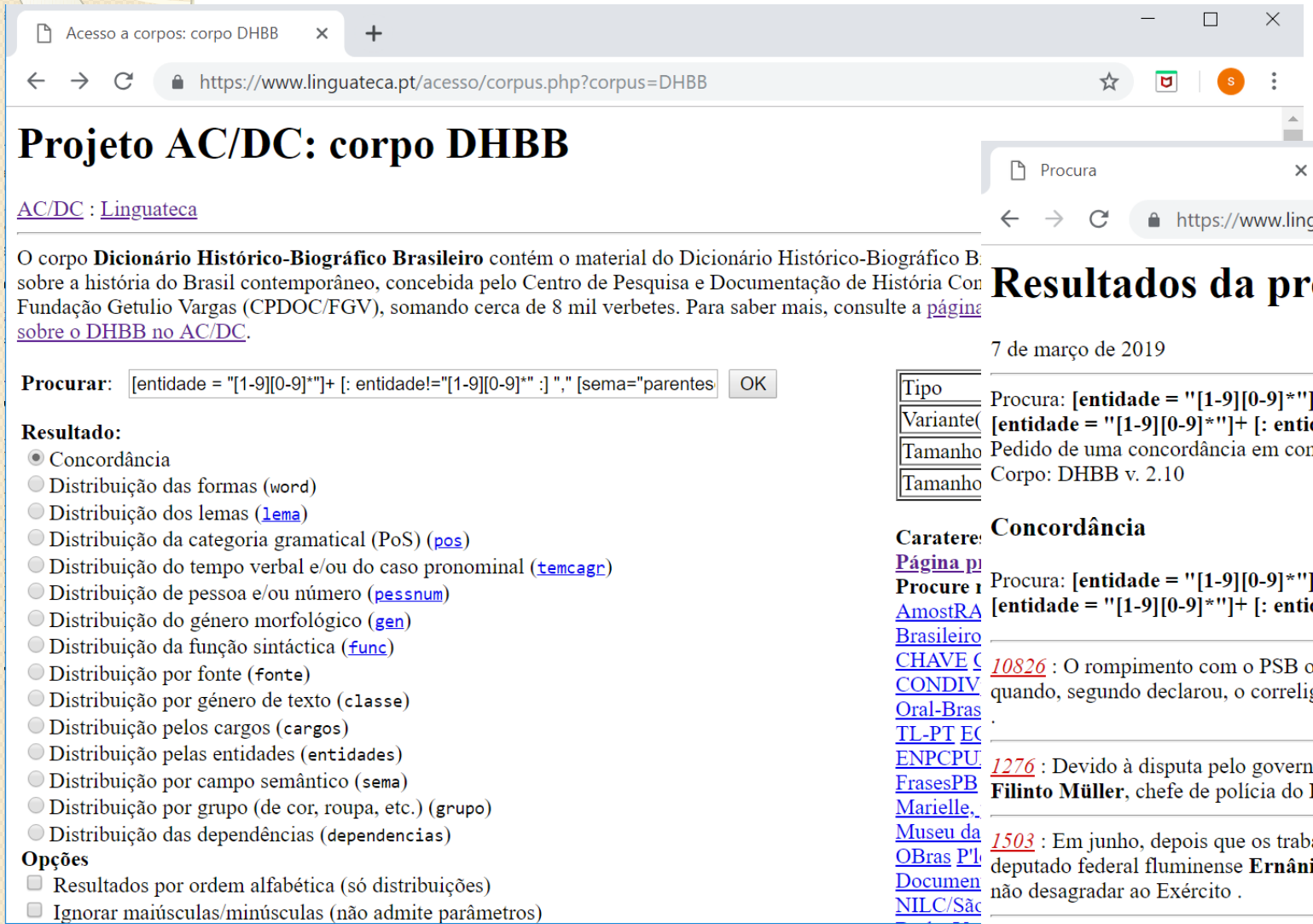
Family connections

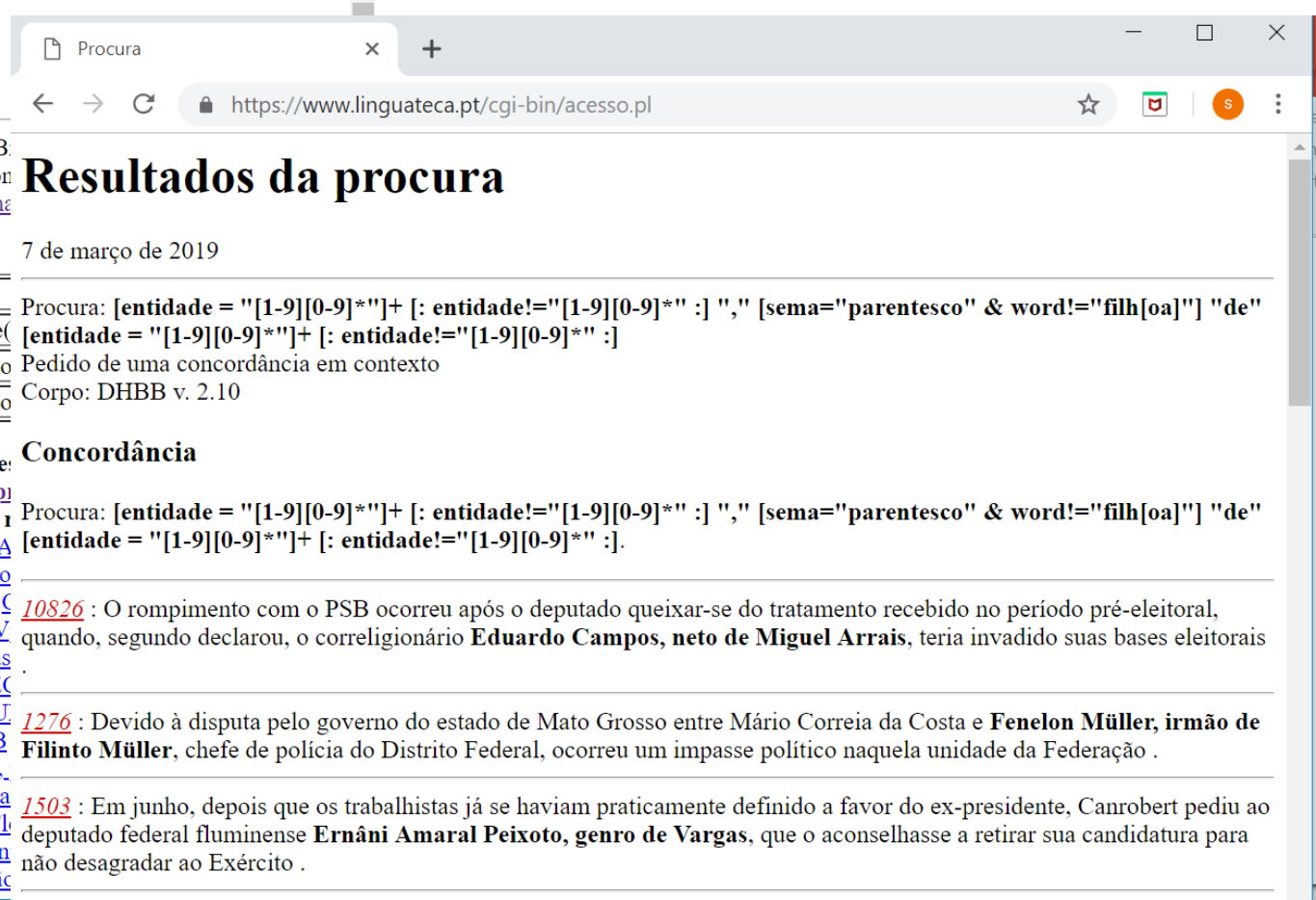
- How to measure, and identify, political dynasties?
- Political families -> who are the few who have controlled Brazil for decades/centuries?

- Semantic annotation with family lexicon;
- Identification of family relationships between politicians who have an entry in DHBB (grounded entities);
- Not all family information mentioned in DHBB is relevant for this task.

Querying the corpus

- using regular expressions, concatenating in a single query lexical, syntactical and semantic information





Family connections – overall picture

filho	10900
pai	1778
irmão	1627
filha	1329
tio	371
primo	345
esposa	298
mãe	274
sobrinho	221
irmã	155
marido	149
avô	135

son → most common family term in Brazilian politics (even after removing the obligatory affiliation cases);
father → second most common, followed by brother

Some distant reading

Organizations and places most mentioned : Law Faculty and Rio (former capital)

Faculdade de Direito

Other distant readings

Most mentioned politicians:

Politician	Occurs
5458 – Getulio Dornelles Vargas	10,181
2412 - João Belchior Marques Goulart	6,602
1418 - Fernando Afonso Collor de Melo	6,214
1072 - Fernando Henrique Cardoso	6,191
3036 - Luís Inácio da Silva	5,706
3807 - Tancredo de Almeida Neves	4,108
4909 - José Ribamar Ferreira de Araújo Costa	3,904
...	...

One can also find in which context these politician are mentioned in the corpus

Other distant readings

Relationships among politicians (sample):

In the entry about	Occurs	No. mentions
Getulio Dornelles Vargas	Osvaldo Euclides de Sousa Aranha	158
Getulio Dornelles Vargas	João Neves da Fontoura	146
Pedro Aurélio de Góis Monteiro	Getulio Dornelles Vargas	91
Osvaldo Euclides de Sousa Aranha	João Alberto Lins de Barros	58
Juscelino Kubitschek de Oliveira	Carlos Frederico Werneck de Lacerda	39
...		

With these data we intend to create a network representation

Other distant readings

- **formal education** of the federal deputies (*deputados federais*) elected by a specific circle, for example, the state of Rio de Janeiro?

```
[cargos=".*depfedRJ.*" & lema="formar.*|licenciar.*|bacharelar.*|
graduar.*|cursar.*|estudar.*"][pos="PRP"]* @[pos="N.*"]
```

10839 : **Bacharelou-se em direito** pela PUC em 1964 .

10853 : Concluiu o ensino médio no Centro Educacional da L
e **graduou-se em marketing** pela Faculdade Estácio de Sá er

10897 : **Formada em pedagogia** pela SUAM (RJ) em 1986,
municipal de promoção social em São João de Meriti no mesr

10903 : De 1987 a 1988 **estudou administração** de empresas
Paulo Apóstolo, mas não concluiu o curso .

10942 : **Bacharelou-se em direito** pela Universidade Federal
em 1991 .

10945 : Professora e funcionária pública, **estudou economia** na Faculdade do Centro Educacional de Niterói (Facen) , mas não concluiu o curso .

"obtained a bachelor's degree in **law**…"

"graduated in **marketing**..."

"graduated in **pedagogy**..."

"studied **administration**..."

"obtained a bachelor's degree in **law**…"

"studied **economics**..."

Distribuição

Houve **63** valores diferentes de **lema**.

ciência	42
direito	28
engenharia	11
medicina	11
economia	7
administração	6
geografia	4
teologia	4
seguida	4
pedagogia	3
história	3
letra	3
engenheiro	3

Problems and limitations

- DHBB is not updated. There is a large team of experts who are writing entries, but it is not real time!
- The writing process has developed for 4 decades: possibly different styles from different DHBB writers
- Linguateca works in the model “human in the loop”, which means that automatic analysis is refined and corrected, but this does not guarantee 100% correctness. There are always errors to be corrected

Future work

- Annotate with other semantic domains that appear relevant to the study of Brazilian politics (temporal information, professions, ethnic background, law charges, etc.)
- Evaluate DHBB annotation with small probes done by close reading
- Apply techniques and methods of visualization of network analysis.

Thank you!

suemi.higuchi@gmail.com