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

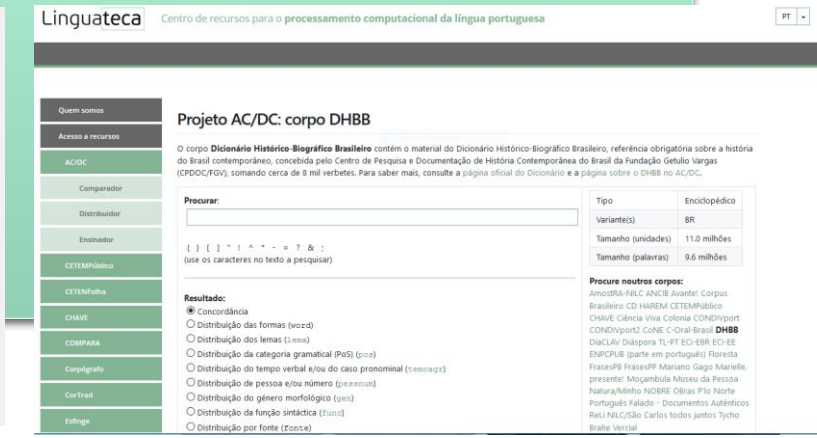
Suemi Higuchi – CPDOC/FGV
Claudia Freitas – PUC Rio
Diana Santos – Linguateca & UiO

PROPOR 2022 - International Conference on the Computational Processing of Portuguese
March 21st-23rd, 2022

The resource

- *Dicionário Histórico-Biográfico Brasileiro* (DHBB) from FGV;
- Encyclopedic work;
- More than 7,500 biographical and thematic entries;
- Domain: contemporary history of Brazil (after 1930)

1984	2001	2010	2018
Print version (4,500 entries)	Print version and cd-rom (6,600 entries)	Online version on the internet (7,500 entries)	Corpus version available at Linguateca



The research

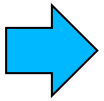
- **Aim**: to create, from the DHBB, an annotated corpus for automatic information extraction's purpose

Motivation: Researches → "Is it possible to extract certain information without having to consult individual entries?"

Examples:

- How is the educational background of political cadres characterized over time?
- How old were Supreme Court ministers when they were nominated?
- Who are the politicians who have family ties to other politicians? Which ties?

Metadata → not enough
Close reading → hard task



Solution: NLP methods and techniques
Distant reading of the DHBB

DHBB overview

(version 7.3)

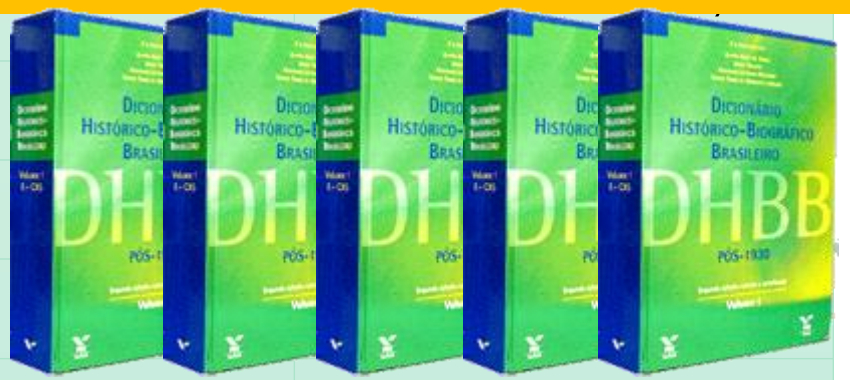
Biographies of:	Occurrences:
Presidents of Brazil	26
Ministers	776
Ministers of the STF	96
Ministers of the STM	118
Senators	627
Federal Deputies	3,835
Military	704
Participants of revolutions/movements	368
Journalists	196
...	...

DHBB overview

(version 7.3)

Biographies of:	Occurrences:
Presidents of Brazil	
Ministers	
Ministers of the STF	
Ministers of the STM	
Senators	
Federal Deputies	
Military	
Participants of revolutions/movements	
Journalists	
...	7,500 pages ...

Entries	7,603
. biographical	6,669
. thematic	934
Tokens	10,079,830
Sentences	314,168



About the DHBB textual structure

- Standardized writing and ‘well-behaved’ text;
- Production of an entry: sentence templates, guidelines for paragraphs sequence, verb tenses to use, etc.

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.
....

- Birth, family information and personal attributes
- Academic background
- Professional activities
- Political trajectory
- Deceased details (if any)
- Spouses and children
- Works produced and publication dates.

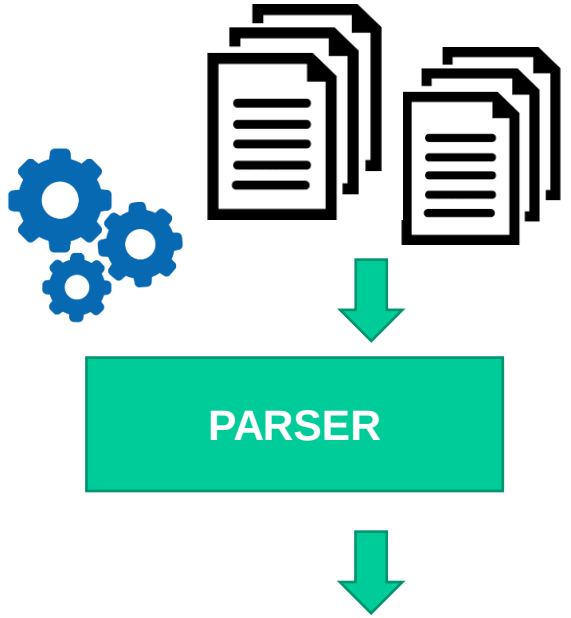
Methodology

- Creation of the corpus
  - Available at https://github.com/cpdoc/dhbb
- Annotation of the corpus
  - Using PALAVRAS parser
- Incorporation into Linguateca (AC/DC format)
  - Available at https://www.linguateca.pt/acesso/corpus.php?corpus=DHBB
- More annotation
  - Processed by AC/DC project
- Addition of semantic layer
- Improvement NER with lists
- Definition of the themes for extraction
- Identification of the set of patterns
- Extraction using the patterns

Methodology

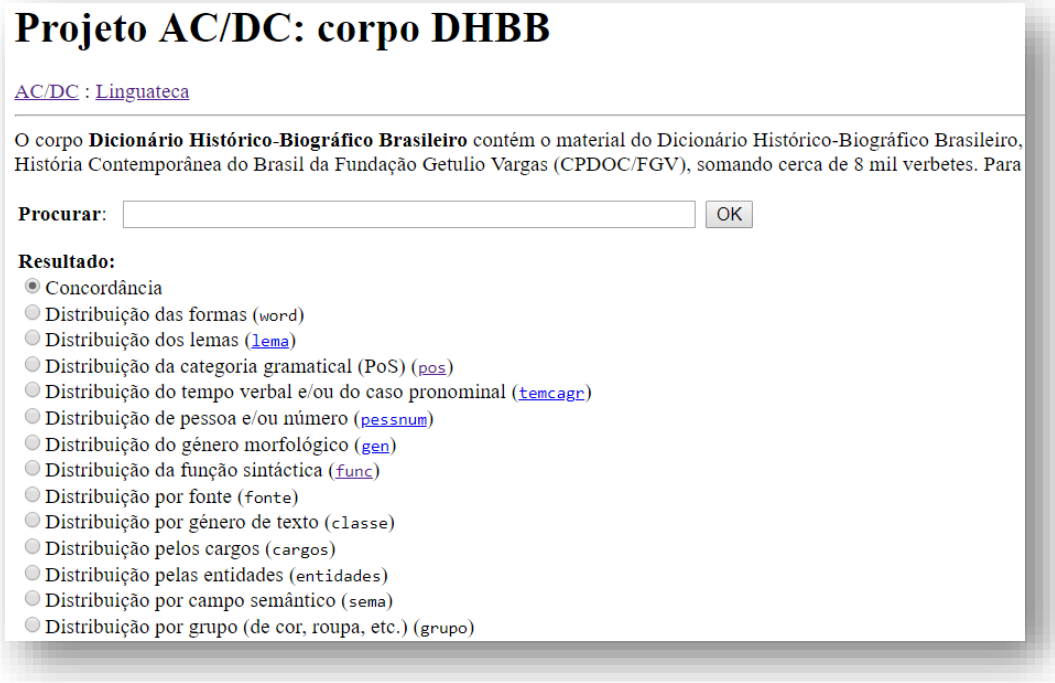
Corpus creation and annotation

each entry = a document containing metadata and text



- Tokenization
- Morphosyntactic and semantic analysis
- Conversion to AC/DC format
- More morphosyntactic and semantic analysis

```
---
titulo: GOULART, João
natureza: biográfico
sexo: m
cargos:
 - dep. fed. RS 1951
 - dep. fed. RS 1952-1953
 - min. Trab. 1953-1954
 - dep. fed. RS 1954
 - vice-pres. Rep. 1956-1961
 - pres. Rep. 1961-1964
---
```



Annotated corpus

Morphosyntactic tags

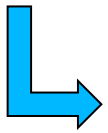
- lemmas
- part of speech (pos tag)
- verb tenses
- singular and plural nouns
- syntactic function

Information provided by metadata:

- full name of the entry holder (*fonte*)
- original identification (*entidade*)
- gender (*sexo*)
- positions held (*cargos*)

Semantic annotations:

- proper names (people, organizations, places, etc.)
- family relationships (lexicon of terms that denote kinship relationship)



Other semantic information can be aggregated = endless enrichment possibilities

Use of rules to improve NER

Correspondence between names referring to the same person:

Lemmas (nicknames, parts of the name, typos...)		DHBB entry name (entidade)
Bolsonaro	=	Jair Messias Bolsonaro
Getulio Vargas	=	Getúlio Dornelles Vargas
Jango	=	João Belchior Marques Goulart
Lula	=	Luis Inácio da Silva

To fix segmentation errors:

Simpatizante	de	Vargas
PoS = noun	PoS = preposition	PoS = proper name

In this case it should only recognize “Vargas”, but it had “Simpatizante” added to its name.

Methodology

The methodology is based on the use of textual patterns, where extraction rules are manually created and rely mainly on lexical-syntactic aspects and semantic annotation of the corpus

Methodology

For each theme:

Compilation of the set of patterns

- observe in a sample of entries how the sentences that bring the desired information are constructed;
- separate as many of them as possible, considering diversity and scope;
- translate these constructions into lexical-syntactic patterns with regular expressions;
- concatenate all expressions.

Extraction

- query the corpus;
- postprocess the results using R;
- specific information is extracted and crossed with metadata.

Extraction

Year of birth

Only one pattern was needed

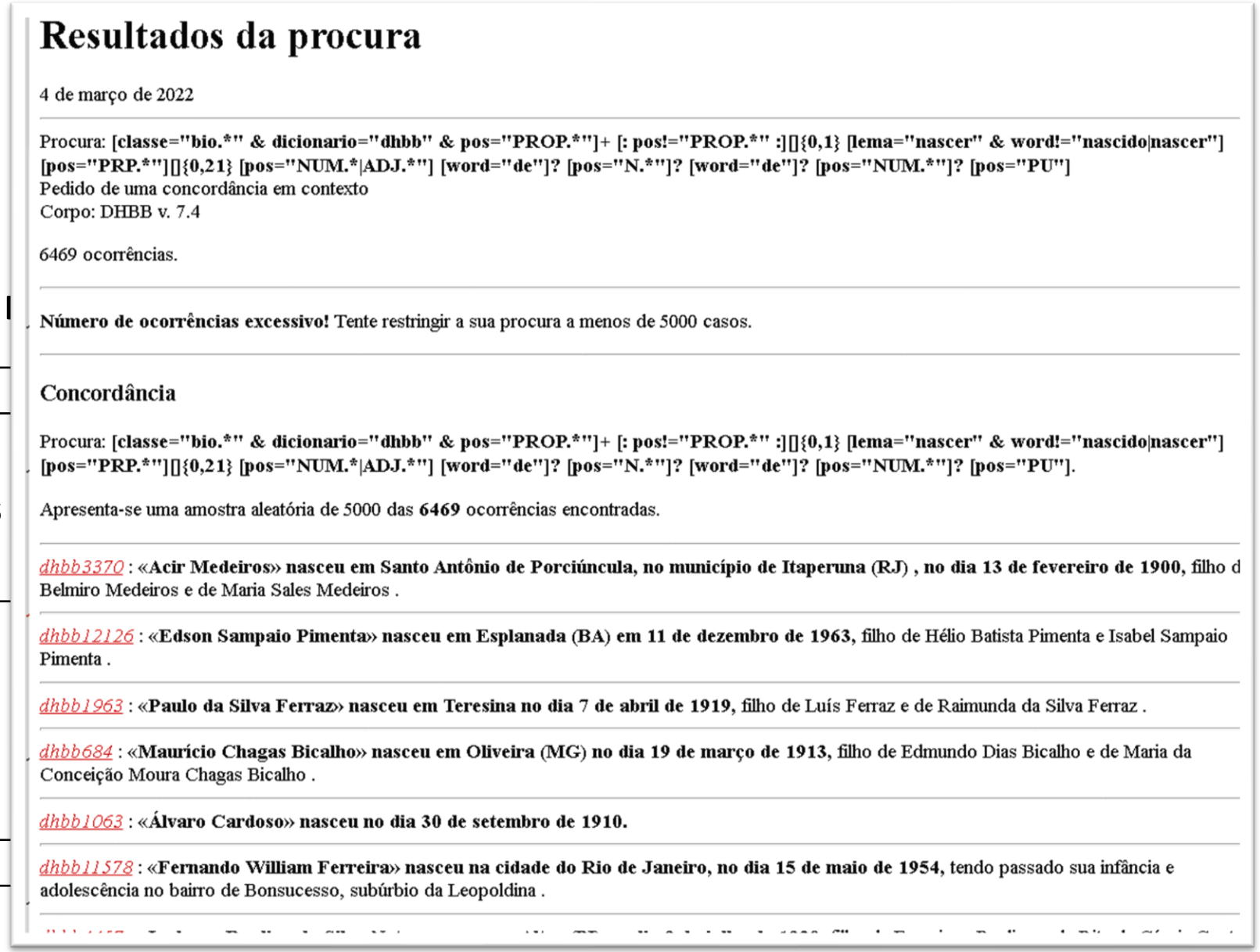
NASCIMENTO	
Examples of sentences	«Moroni Bing Torgan» nasceu em Porto Alegre, no dia 10 de junho de 1956.
	«Álvaro Francisco de Sousa» nasceu no dia 28 de fevereiro de 1903.
Pattern	[classe="bio.*" & dicionario="dhbb" & pos="PROP.*"]+ [: pos!="PROP.*" :][]{0,1} [lema="nascer" & word!="nascido|nascer"] [pos="PRP.*"][]{0,21} [pos="NUM.*|ADJ.*"] [word="de"]? [pos="N.*"]? [word="de"]? [pos="NUM.*"]? [pos="PU"]
Occurrences	6,469

Extraction

Year of birth

Only one pattern was

NASCIMENTO
Examples of sentences
Pattern
Occurrences



Resultados da procura

4 de março de 2022

Procura: **[classe="bio.*" & dicionario="dhbb" & pos="PROP.*"]+ [: pos!="PROP.*" :][]{0,1} [lema="nascer" & word!="nascido|nascer"] [pos="PRP.*"][]{0,21} [pos="NUM.*|ADJ.*"] [word="de"]? [pos="N.*"]? [word="de"]? [pos="NUM.*"]? [pos="PU"]**
Pedido de uma concordância em contexto
Corpo: DHBB v. 7.4

6469 ocorrências.

Número de ocorrências excessivo! Tente restringir a sua procura a menos de 5000 casos.

Concordância

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

Apresenta-se uma amostra aleatória de 5000 das **6469** ocorrências encontradas.

dhbb3370 : «**Acir Medeiros**» **nasceu em Santo Antônio de Porciúncula, no município de Itaperuna (RJ) , no dia 13 de fevereiro de 1900,** filho d Belmiro Medeiros e de Maria Sales Medeiros .

dhbb12126 : «**Edson Sampaio Pimenta**» **nasceu em Esplanada (BA) em 11 de dezembro de 1963,** filho de Hélio Batista Pimenta e Isabel Sampaio Pimenta .

dhbb1963 : «**Paulo da Silva Ferraz**» **nasceu em Teresina no dia 7 de abril de 1919,** filho de Luís Ferraz e de Raimunda da Silva Ferraz .

dhbb684 : «**Maurício Chagas Bicalho**» **nasceu em Oliveira (MG) no dia 19 de março de 1913,** filho de Edmundo Dias Bicalho e de Maria da Conceição Moura Chagas Bicalho .

dhbb1063 : «**Álvaro Cardoso**» **nasceu no dia 30 de setembro de 1910.**

dhbb11578 : «**Fernando William Ferreira**» **nasceu na cidade do Rio de Janeiro, no dia 15 de maio de 1954,** tendo passado sua infância e adolescência no bairro de Bonsucesso, subúrbio da Leopoldina .

Extraction themes

- The age of entrance in Brazilian politics;
- The academic background of Brazilian politicians;
- family ties among the political elites.

Extraction evaluation

Year of birth

All occurrences were manually checked

Occurrences	Retrieval of sentences with birth information		
	All DHBB biographical entries		
Total entries	6.745 (100%)		
Retrieved entries	6.455 95,7%	VP	6.444
		FP	11
Unrecovered entries	290 4,3%	VN	235
		FN	54
Precision	.99		
Recall	.99		
F-Measure	.99		

Extraction evaluation

Education background/academic training

Example sentences	se graduou em ciências sociais na Universidade de São Paulo (USP).
	diplomou-se como engenheiro naval em 1964, na Escola Nacional de Engenharia.
Example pattern	[word="se"]? [dicionario="dhbb" & classe="biográfico" & lema="especializar.*|diplomar.*|bacharelar.*|graduar.*|doutorar.*" & word!= "especializad.?s|especializada"] []* "\."
Occurrences of this pattern	2.890

11 pattern expressions
10.565 occurrencies
5,627 bio entries (83% total)
All manually checked

Total occurrences	**10.565 (100%)**
Valid occurrences	10.470 (99,1%)
Invalid occurrences	95(0,9%)

It is not possible to know how many and which occurrences were not recovered throughout the corpus, so the assessment does not include a measure for recall.

Extraction evaluation

Family ties

Example sentences	Seu filho, Fernando Ramos de Alencar, sobressaiu-se na carreira diplomática diplomou-se como engenheiro naval em 1964, na Escola Nacional de Engenharia.
Example pattern	[word="se"]? [dicionario="dhbb" & classe="biográfico" & lema="especializar.*|diplomar.*|bacharelar.*|graduar.*|doutorar.*" & word!= "especializad.?s|especializada"] []* "\."
Occurrences of this pattern	2.890

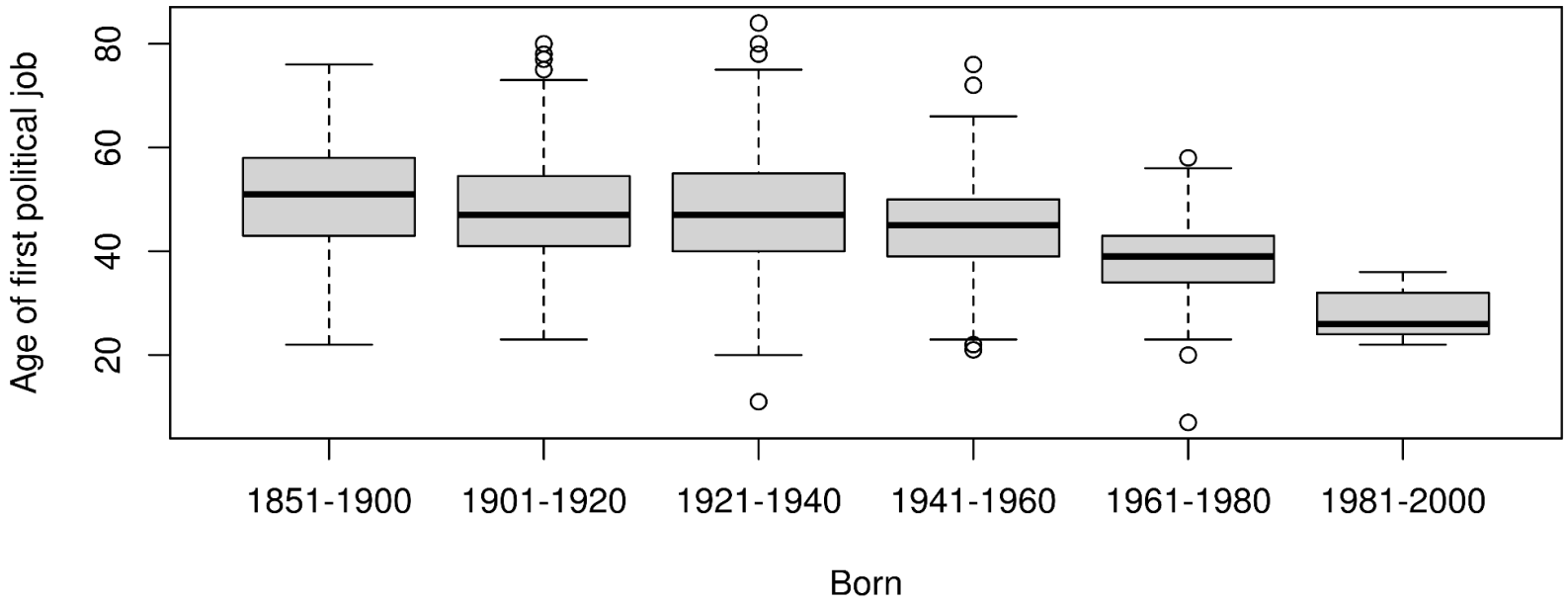
33 pattern expressions
6,220 occurrencies

A sample of 198 random cases were manually checked: **average precision of .59**

→ our interest was not to know whether family ties were correctly identified using these patterns (and most of them were), but rather to measure the proportion of family relationships among politicians that can be found in the corpus.

Some distant readings from DHBB:

The age of starting public careers, per generation



- Those aspiring to political careers are getting in the public service steadily younger
  - The difference in the average between the so-called generation 1 (born before 1900) and generation 6 (born after the 1980s) falls almost by half
  - it was more common to start a public career around the age of 50 years old hundred years ago,
  - this changed until it reached an average of 27 years old in the last generation

Some distant readings from DHBB:

The education background of the politicians

Areas of training most frequent among the biographees, throughout the generations

Area	ALL	undated entries	b. 1900 (geração 1)	1901-1920 (geração 2)	1921-1940 (geração 3)	1941-1960 (geração 4)	1961-1980 (geração 5)	1981-2000 (geração 6)
Direito	2.914	59	618	711	829	559	133	5
Militar	854	18	316	326	162	27	4	1
Engenharia	733	16	157	135	189	206	27	3
Medicina	672	19	157	163	122	195	15	1
Administração	516	11	7	35	132	256	71	4
Economia	517	11	10	73	132	258	32	1
Filosofia	267	7	26	63	97	64	10	-
Contabilidade	186	7	4	42	77	52	4	-
Letras	179	2	30	35	46	50	16	-
C. Sociais	137	2	9	17	40	55	13	1

- among the 48 found areas, in the overall reckoning law school appears preponderant in all generations, followed by military training, with just one third of the first;
- decline in military training from the second to the third generation;
- civilian training was replacing the military career as the most suitable way to reach important political positions.

Some distant readings from DHBB:

Family ties in politics

Distribution of family ties for Presidents and Senators

Biographies with family ties/ total number of biographies	**Born**:	**Identification of at least one family tie with another biographee / total of biographees**			
		Men	Women	ALL	%
Presidents 13 /26 50 %	NO DATE	0 /2	-	0 /2	0 %
	Before 1900	6 /12	-	6 /12	50 %
	1901 – 1920	5 /8	-	5 /8	62,5 %
	1921 – 1940	2 /2	-	2 /2	100 %
	1941 – 1960	0 /2	-	0 /2	0 %
	1961 – 1980	-	-	-	0 %
	1981 – 2000	-	-	-	0 %
Senators 260 /742 35 %	NO DATE	5 /17	0 /1	5 /18	27,7 %
	Before 1900	89 /169	-	89 /169	52,6 %
	1901 – 1920	72 /193	-	72 /193	37,3 %
	1921 – 1940	56 /183	0 /5	56 /188	29,8 %
	1941 – 1960	31 /139	4 /12	35 /151	23,2 %
	1961 – 1980	3 /17	0 /6	3 /23	13,0 %
	1981 - 2000	-	-	-	-

Some distant readings from DHBB:

Family ties in politics

Presidents and senators are the politicians who most appear with family ties, 50% and 35% respectively, this being most perceived in the first generations (in the present version of DHBB). Ministers and deputies follow with 18%, keeping a stable average over generations.

Distribution

Biographies with family ties/ total number of biographies	Born:	Identification of at least one family tie with another biographee / total of biographees			
		Men	Women	ALL	%
Presidents 13 /26 50 %	NO DATE	0 /2	-	0 /2	0 %
	Before 1900	6 /12	-	6 /12	50 %
	1901 – 1920	5 /8	-	5 /8	62,5 %
	1921 – 1940	2 /2	-	2 /2	100 %
	1941 – 1960	0 /2	-	0 /2	0 %
	1961 – 1980	-	-	-	0 %
	1981 – 2000	-	-	-	0 %
Senators 260 /742 35 %	NO DATE	5 /17	0 /1	5 /18	27,7 %
	Before 1900	89 /169	-	89 /169	52,6 %
	1901 – 1920	72 /193	-	72 /193	37,3 %
	1921 – 1940	56 /183	0 /5	56 /188	29,8 %
	1941 – 1960	31 /139	4 /12	35 /151	23,2 %
	1961 – 1980	3 /17	0 /6	3 /23	13,0 %
	1981 - 2000	-	-	-	-

Final considerations

Some challenges and issues:

- how to find a balance between a sufficient number of patterns and a good enough coverage?
- the manual work required to create the expressions;
- Languages natural expressiveness.

Final considerations

Main contributions:

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

Thank you