Text Mining for History: First Steps on Building a Large Dataset

Suemi Higuchi[1,2], Cláudia Freitas[1], Bruno Cuconato[3], Alexandre Rademaker[3,4]
alexrad@br.ibm.com
PUC-RJ[1], FGV/CPDOC[2], FGV/EMAp[3], IBM Research[4]

Motivation

1. The mining strategy is linguistically motivated: certain semantic relations have a linguistic realization; the inclusion of linguistic metadata such as PoS, lemma, and syntactic information in the corpus is essential.

2. "Which politicians were born before the 1960s, had military training and held a position in the Executive Branch?" Vast amount of knowledge spread around the entries in a non-linear way.

3. DHBB entries are written in encyclopedic style, and this "novelty" can be a challenge for automatic parsers.

4. Appositives are productive for text mining: noun phrase co-reference, information extraction etc. Induce many semantic relations: `ident`, `role` and `link-fam` all appear on the text.

DHBB

1. 12M tokens in about 300K sentences. It is written by historians and published by FGV/CPDOC. 8K entries with information ranging from the life and career trajectories of individuals to the relationships between the characters and events in Brazil.

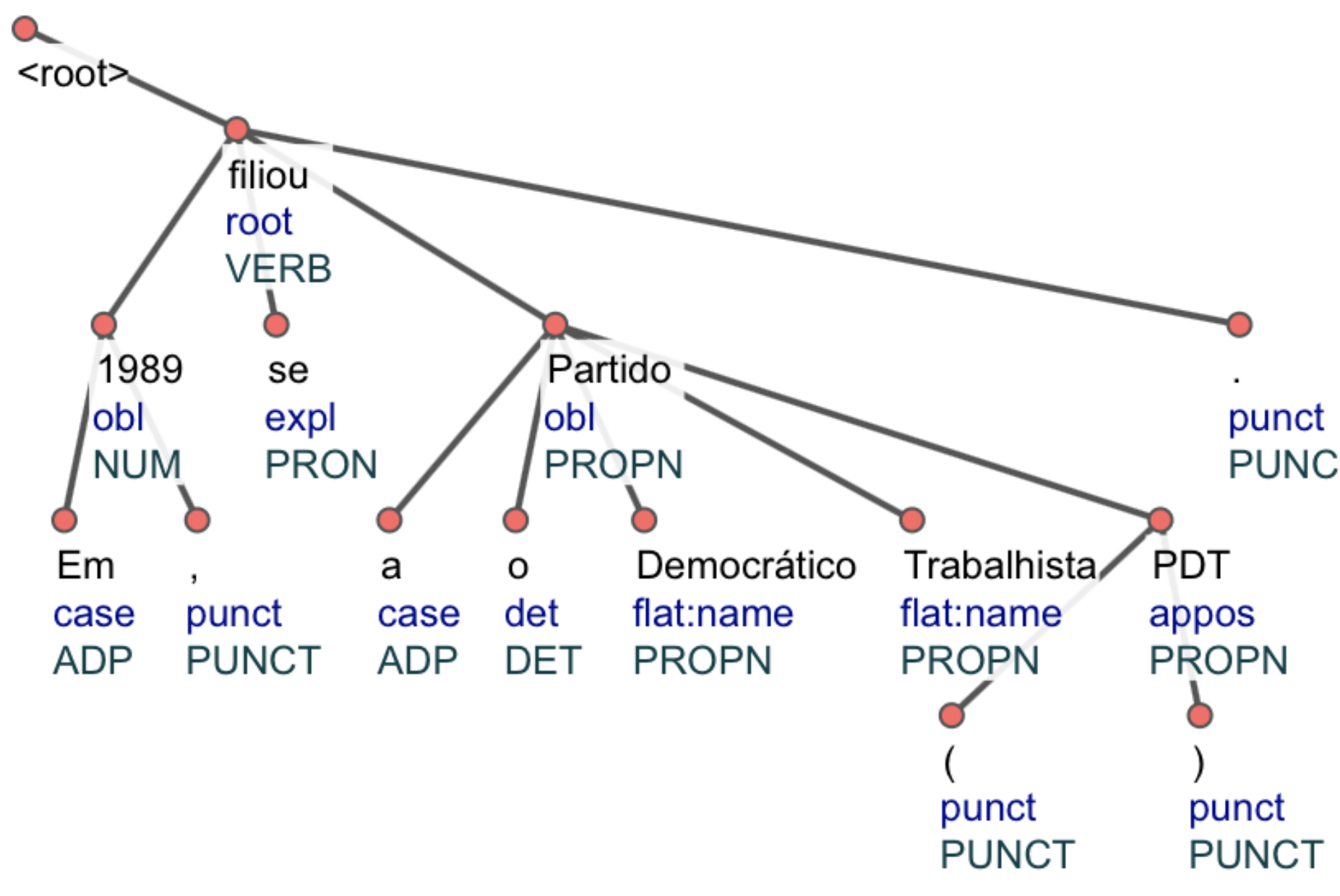
2. Entries are in text files using a lightweight human-readable markup syntax: YAML. First version from 1984.

3. The data is freely available at `https://github.com/cpdoc` and `https://cpdoc.github.io/dhbb`.

Corpus Preparation

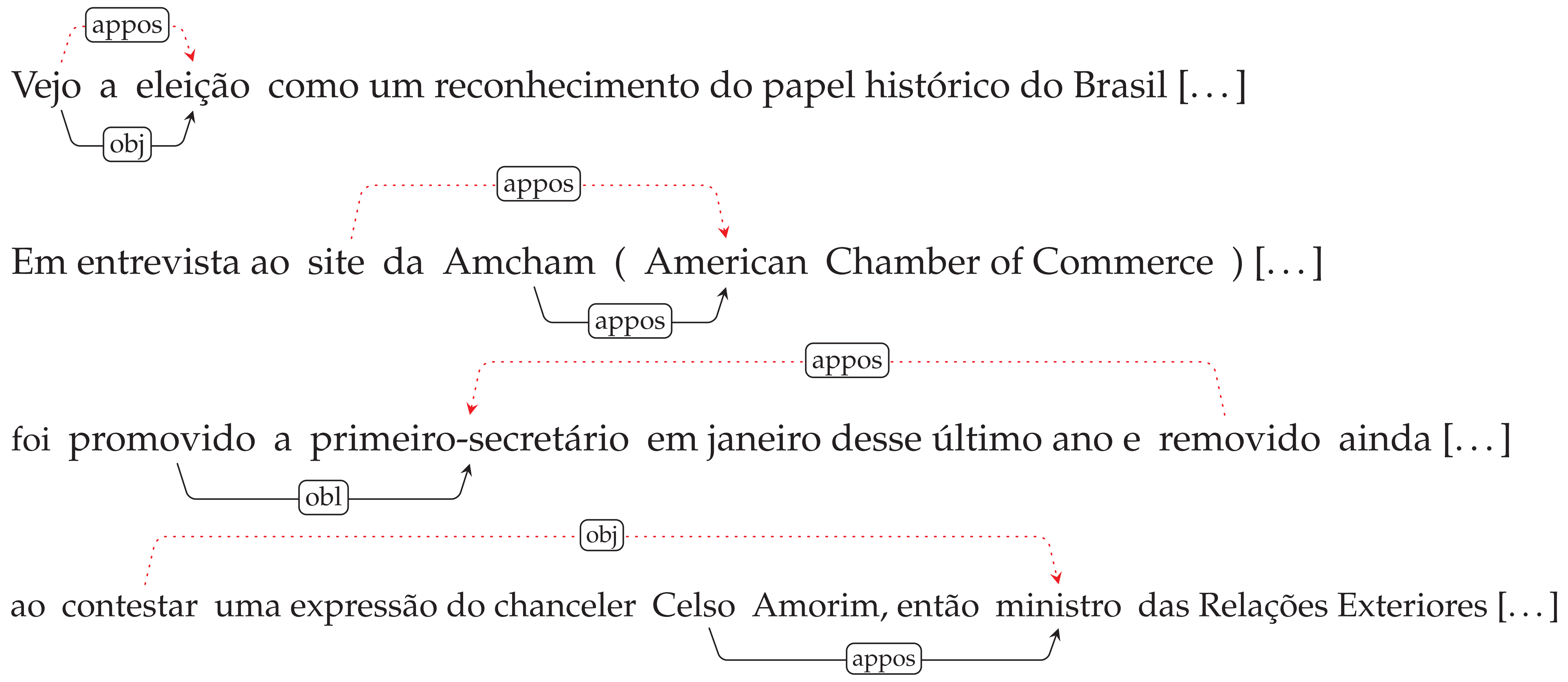
PALAVRAS and UDPipe (Bosque Corpus):

```
Em [em] <*> PRP @ADVL> #1->4
1989 [1989] <card> <date> <year> NUM M/F S @P< #2->1
, #3->0
filiou- [filiar] <fmc> <hyfen> <vH> <mv> <v-pron> V PS 3S IND VFIN @FS-STA #4->0
se [se] <refl> PERS M/F 3S ACC @<ACC #5->4
a [a] <sam-> PRP @<PIV #6->4
o [o] <-sam> <artd> DET M S @>N #7->8
Partido Democrático Trabalhista [Partido=Democrático=Trabalhista] <party> <*> PROP M S @P< #8->6
( #9->0
PDT [PDT] <party> <*> <np-close> PROP M S @APP #10->8
) #11->0
. #12->0
```



Manual entity relations annotations for the sample, and an entity lexicon built semi-automatically from lexical-syntactical patterns, taking advantage of the highly predictable written style of the DHBB.

Errors



Vejo a eleição como um reconhecimento do papel histórico do Brasil [. . .]

Em entrevista ao site da Amcham (American Chamber of Commerce) [. . .]

foi promovido a primeiro-secretário em janeiro desse último ano e removido ainda [. . .]

ao contestar uma expressão do chanceler Celso Amorim, então ministro das Relações Exteriores [. . .]

Entities and Relations

Entities: PER, ORG, POL (pol. formulation), EVN, LOC, DOC, TME.

ident (correference) *Partido dos Trabalhadores* and *PT*

role *Alberto Coelho* and *president*

ploc (local) *port of Alcantara* and *in Lisbon*

part *Porto Seguro* and *BA*

date *promulgation of Nova Carta* and *18/9/1946*

link-inst *Vandilson Costa* and *from Partido Comunista do Brasil*

link-fam (family relation) *Nilo Augusto* and *son of Gercino Coelho and Eunice Coelho.*

link-pers (personal relation) *Orígenes Lessa* and *friend of his brother Fúlvio*

attrib (attribute) *João Abdalla and Amélia Abdalla* and *of Arab origin*

participant *Getulio Vargas* and *in the Revolution of 1930*

context *XXXVIII ministerial meeting* and *of General Agreement on Tariff and Trade*

Segmentation

1. Names such 'Ministério das Minas e Energia' and 'José Afonso de Melo' are hard.

2. Lexicon: metadata for person names and patterns (AntConc) to extract names of organizations.

3. 129,456 person names (8,642 unique), 48% of the whole list. Organizations 99,384 names (3,537 unique) occurring in the corpus, represents 97% of the lexicon.

4. 790 names from the lexicons were found in the golden, correct 460 (58%) tokens, only 30 (0.03%) of the affected tokens had an appos relation.

The Experiment

1. 35 entries: 38,554 tokens in 1,115 sentences, 472 (>1 appos), 796 appos with 10 types relations.

2. (i) revising the segmentation of the names; (ii) manually identifying the induced semantic relationship; (iii) assessment the quality of the parsers (PAL vs UDPipe); and (iv) assessment the impact of NE domain lexicon in PROPN segmentation.

3. $(N - Head, appos, N - Head)$ could be trivially analyzed and **abnormal noun phrases** indicate **parser mistake**.

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

Analysis

AllCorrect correct idt of the args of the relation and correct idt of appositive.

ErrDepRel correct idt of the args of the relation but incorrect idt of appositive.

ErrHead incorrect idt of the args of the relation but correct idt of appositive.

FullErr incorrect idt of args and relation.

MissingAppos appositive relation was not detected.

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