Identifying family ties among politicians
Challenges of information extraction evaluation

Diana Santos[1][0000−0002−3108−7706]*, Suemi Higuchi[2][[0000−0002−6255−3781], and Cláudia Freitas[3][0000−0001−6807−8558]

[1] Linguateca & University of Oslo, Norway
d.s.m.santos@ilos.uio.no
[2] FGV, Brazil
Suemi.Higuchi@fgv.br
[3] Linguateca & PUC-Rio, Brazil
claudiafreitas@puc-rio.br

Abstract. We discuss several challenges of evaluating information extraction patterns, using the DHBB corpus, a public resource for the Dicionário Histórico-Biográfico Brasileiro. Our goal is to stress both the limitations and the advantages of using a corpus-based approach for the task of identifying political families in Brazilian society.

Keywords: Evaluation · Information extraction · Brazilian politics

1 Family ties in Brazilian politics

It is often mentioned that in Brazil family ties matter a lot for success in politics [12, 7]. However, this is not easy to measure and therefore confirm. But given the availability of the DHBB corpus, we decided to extract all family relationships there mentioned, and assess whether they concerned family relationships among politicians.

This can be considered a kind of distant reading for History [2, 10], and it highlighted the need to be very concrete as to what exactly one is evaluating.

We begin by explaining how we annotated family ties in the DHBB corpus, then how we annotated that a particular name was already a biographee in DHBB, and then how the family relations were extracted. Then we discuss how to evaluate the result, and show that there are several ways one can evaluate the resulting concordances.

In our understanding, a politician is someone who is invested in his or her position through election, nomination or designation, usually members of the executive and legislative branches[4]. Positions that serve merely for bureaucratic

[4] https://corregedorias.gov.br/assuntos/perguntas-frequentes/agentes-publicos-e-agentes-politicos.

functions, such as technical advisers and consultants, whether executive, legislative, judiciary branches or military, are generally not considered politicians, although they are involved in government decision-making processes.[8]

2 Annotating family in the AC/DC

DHBB belongs to the AC/DC family of corpora, a project designed to make available, and searchable, large corpora on the Web [9]. Because we were especially interested in family ties in the context of Brazilian politics, we added them as one of the semantic fields available in AC/DC, something that may be relevant as well for other kinds of text, as the current organization of DIP [11] for extracting chracters and their family ties demonstrates[5].

3 Grounding biographees

Contrary to the family semantic domain, there is information that only makes sense for DHBB, namely the unification of several distinct names as corresponding to a particular biographee (Lula, Lula da Silva, Luís Inácio Lula da Silva, for example). In fact, each politician who has an entry in DHBB receives an unique identifier (stored in the `id` field), and during creation of the corpus we tried to unify the several different ways of referring to the same person, using manually constructed rules, as described in [5].

4 Obtaining extraction patterns

By looking at ten entries whose holders are known to have many family ties with other politicians, the second author devised a set of patterns, divided in five groups (not necessarily mutually exclusive), as detailed in [3, page 119].

1. Relations between the entry politician and other people biographed in DHBB
2. Relations between the entry politician, using the possessive pronoun, assuming that it refers to the biographee, and named politicians.
3. Relations between the entry politician and another non-named politician
4. Relations between two politicians biographed in DHBB, none of them the biographee (This gave us 35 cases, all of them correct.)
5. General family relations described in DHBB between names

Sentences of each group are listed below, preceded by type:

- (1) Paulo Maluf, seu padrinho, (...)
- (2) **Sua esposa era filha de João Alves de Sousa**, militar, tenente-coronel e chefe político em Patrocínio (MG) .
- (3) Seu pai foi eleito deputado constituinte (...)

[5] More information on the categories can be found in `https://www.linguateca.pt/Gramateca/Familia.html`.

- (5) (...) os ex-pessedistas, liderados por **Crispim Jaques Bias Fortes, secretário de Obras Públicas, filho do ex-governador** José Francisco Bias Fortes (...)

Note that the second example illustrates an indirect relationship between politicians: the biographee is son-in-law of JAS (his wife is daughter of JAS).

5 Evaluating the results

Since the patterns yielded a large number of results, we obtained a sample per kind of pattern, and manually evaluated those 198 cases. We soon noticed that evaluation could be done according to the following criteria:

- did the patterns find valid family relations? (criterion 1)
- did the patterns find family relations between politicians? (crit 2)
- did the patterns find family relations between politicians which were possible to identify in DHBB? (crit 3)

In addition, one could take a strict evaluation of the patterns (and no text around would be included).

- did the patterns extract family relations between politicians? (crit 4)
- did the patterns extract family relations between politicians which were possible to identify in DHBB? (crit 5) (only first names are not enough)

See, for example, the following cases:

- (...) no entanto derrotada dentro do partido que optou por **Gleisi Hoffmann, esposa do ministro** do Planejamento Paulo Bernardo (GH, wife of the minister)
- No mês seguinte, foi acusado de envolvimento na morte de **Severino Alves de Lacerda, filho do ex-prefeito** do município paraibano de Aguiar . (SAL, son of the ex-mayor)

The first case allow us to find in the whole sentence a relationship between two politicians (Gleisi Hoffman and Paulo Bernardo), but not in the extracted pattern. (it yelds YES for the three first criteria). The second, although it identifies a relationship between two politicians, does not allow us to identify the second politician, and thus yields YES only for for the first, second and fourth criteria.

In Table 1 we show the results of this fivefold evaluation [1]. This shows that a lot of decisions have to be agreed upon, and that evaluation depends on exactly what is one interested in. The results of case 3 were reported in [4].

What we would like to stress here is that all these numbers are appropriate, but evaluate different things. While the first only looks at the precision of family relations in encyclopedic text, the second and fourth measure politician family links, and the third and fifth measure the capability of extracting links among named politicians. The difference between the 2nd and 3rd vs the 4th and 5th deal with the amount of text to be processed. Obviously, the pattern itself is much easier to employ to get triples of the form A-family link-B, that can for example be depicted in a graph.

Table 1. Evaluating the extraction in five different ways

Pattern set	crit 1	crit 2	crit 3	crit 4	crit 5
1	1	.48	.48	.32	.30
2	.98	.64	.64	.46	.44
3	1	.88	.88	.62	.60
5	.96	.44	.42	.24	.18
Total	.985	.605	.595	.405	.375

6 New extraction patterns

While analysing the cases obtained, it became clear that the patterns themselves could be significantly improved if we took into consideration the (main) political positions. While reliably annotating the DHBB with all appropriate political positions is not yet performed, we created a short list, improved the rules that mentioned a noun to become a "political position noun", and got new results.

Table 2. Improving the queries with political positions

Pattern set	Before	After
1	3753	2225
3	640	78
5	624	338

We were able to get down from 5,017 cases to 2,641 cases, which is almost a half, as is shown in Table 2. Also, the patterns themselves became more reliable, in the sense of yielding the position of the family member much more frequently.

This is confirmed by a new random set which was again humanly reviewed, see Table 3, and which can be inspected in [6]. Interestingly, most politicians identified had a name in our sample, which casts doubt on the need to separate politicians from identifiable politicians... and shows that the DHBB authors were very careful to name the people mentioned.

Table 3. Evaluating the extraction in five different ways with the new patterns [6]

	crit 1	crit 2	crit 3	crit 4	crit 5
All	.985	.765	.765	.705	.690

Even though the exercise reported here may seem to exhaust the evaluation possibilities, several interesting issues remain to be solved, such as: full names vs. first names only, concordances where more than one family relationship was present, and the automatic recovery of the possessive pronoun's referent. And finally, the common presence in DHBB of family members with power in Brazil, although not politicians in our sense.

References

1. Avaliação dos antigos padrões com 5 critérios. Tech. rep., Linguateca (5 January 2022), https://www.linguateca.pt/acesso/dhbb/AvaliacaoCincoCriterios5jan2022.html
2. Fortes, A., Alvim, L.G.M.: Evidências, códigos e classificações: o ofício do historiador e o mundo digital. Esboços: histórias em contextos globais **27**(45), 207–227 (2020)
3. Higuchi, S.: Extração automática de informações: uma leitura distante do Dicionário Histórico-Biográfico Brasileiro (DHBB). Ph.D. thesis, PUC-Rio, Rio de Janeiro (May 2021), http://www.linguateca.pt/documentos/TeseSuemiHiguchi2021.pdf
4. Higuchi, S., Freitas, C., Santos, D.: Automatic information extraction: a distant reading of the Brazilian Historical-Biographical Dictionary. In: PROPOR 2022. Springer (March 2022)
5. Higuchi, S., Santos, D., Freitas, C., Rademaker, A.: Distant reading Brazilian politics. In: Proceedings of 4th Conference of The Association Digital Humanities in the Nordic Countries (Copenhagen, March 6-8 2019) (March 2019), https://www.linguateca.pt/Diana/download/aprDHN2019.pdf
6. Avaliação dos novos padrões com 5 critérios. Tech. rep., Linguateca (5 January 2022), https://www.linguateca.pt/acesso/dhbb/AvaliacaoNovosPadroes5jan2022.html
7. Oliveira, R.C.d., Goulart, M.H.H.S., Vanali, A.C., Monteiro, J.M.: Família, parentesco, instituições e poder no Brasil: retomada e atualização de uma agenda de pesquisa. Revista Brasileira de Sociologia **5**(11), 165–198 (2017), https://dialnet.unirioja.es/descarga/articulo/6227086.pdf
8. Cargos políticos no dicionário histórico-biográfico brasileiro. Tech. rep., Linguateca (5 January 2022), https://www.linguateca.pt/acesso/dhbb/AvaliacaoCincoCriterios5jan2022.html
9. Santos, D.: Corpora at Linguateca: Vision and roads taken. In: Berber Sardinha, T., Ferreira, T.L.S.B. (eds.) Working with Portuguese Corpora, pp. 219–236. Bloomsbury (2014)
10. Santos, D.: Humanidades Digitais e História: algumas observações (16 December 2021), https://www.linguateca.pt/Diana/download/SantosPIDH.pdf
11. Santos, D., Willrich, R., Langfeldt, M., de Moraes, R.G., Mota, C., Pires, E., Schumacher, R., Pereira, P.S.: Identifying literary characters in Portuguese: Challenges of an international shared task. In: PROPOR 2022. Springer (March 2022)
12. Schoenster, L.: Clãs políticos seguem dominando Congresso na próxima legislatura. Tech. rep., Transparência Brasil (nov 2014), https://www.transparencia.org.br/downloads/publicacoes/Cl%C3%A3s%20pol%C3%ADticos%20seguem%20dominando%20Congresso%20na%20pr%C3%B3xima%20legislatura.pdf