Experiments in distant reading... in Portuguese

Diana Santos

Linguateca & University of Oslo
d.s.m.santos@ilos.uio.no

Image from https://bokogbibliotek.no/debatt/heime-pa-biblioteket-4685/

1st Int. Conf. on Data & Digital Humanities, 10 March 2023

Distant reading (my definition)

- Using digital tools to study a collection of texts which have been created for human reading, with the goal to make sense of large numbers of those texts, with modern techiques.
- Two main kinds of texts there has been distant reading on:
  - informative/scientific texts, for which in fact topic modelling was suggested by David Blei
  - literary texts, for which the term "distant reading" was coined by Franco Moretti
- but then a lot of other kinds of texts, like historical texts, newswire, encyclopedias, etc. have been distantly read

Questions discussed in the present talk

- Is it possible to understand/predict literary school? Have a bird's eye of author style?
- What can we say about literary characters in Portuguese novels?
- How are characters depicted? How are they linked in family?
- Places in Portuguese literature and what they tell us (places mentioned in Camilo's works, Santos & Alves, 2023)
- Can we find "inner life" in literature?
- Brazilian politicians as a population

Other subjects not covered in the talk

- Emotions
- Reference to the human body and to clothing
- Health
- Reporting
- Named entities

See Santos et al. (2020) – in Portuguese – for a broader description of the beginnings of distant reading in Portuguese.

Acknowledgements

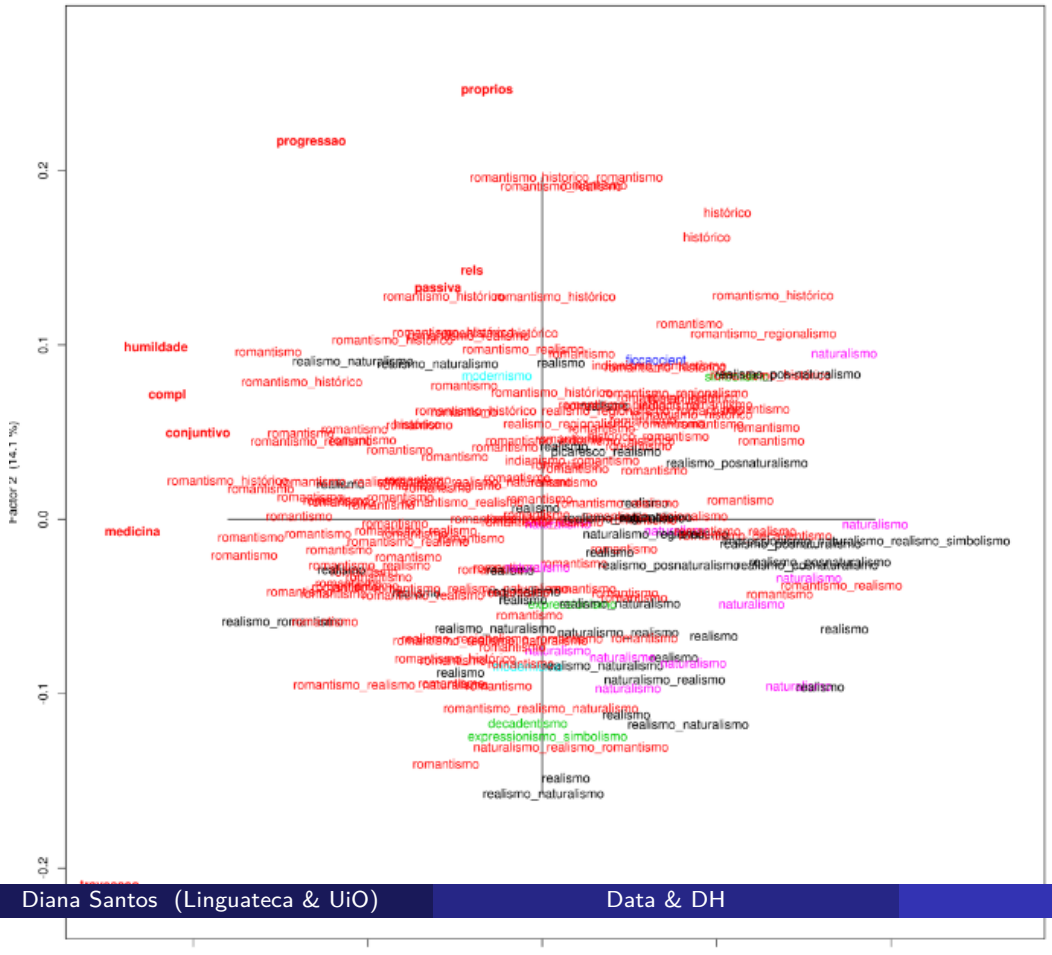
Everything I am presenting here is common work with these colleagues:

- Alberto Simões
- Cláudia Freitas
- Cristina Mota
- Daniel Alves
- Eckhard Bick
- Emanoel Pires
- João Marques Lopes
- Marcia Langfeldt
- Rebeca Schumacher
- Roberto Willrich
- Suemi Higuchi

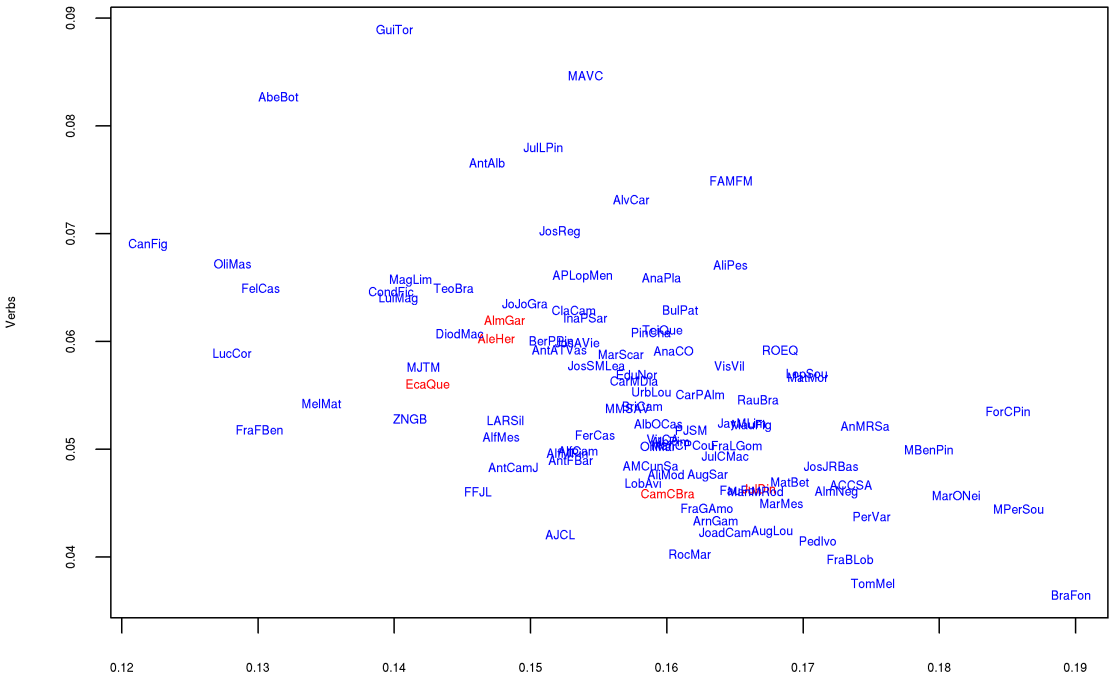
and others

http://folcloredeportugal.blogspot.com/2010/

Does literary school correlate with linguistic features?



Style differences among Portuguese authors in 196 novels (1840-1949)



DIP: *Desafio de Identificação de Personagens*

The character identification challenge...
`https://www.linguateca.pt/DIP`

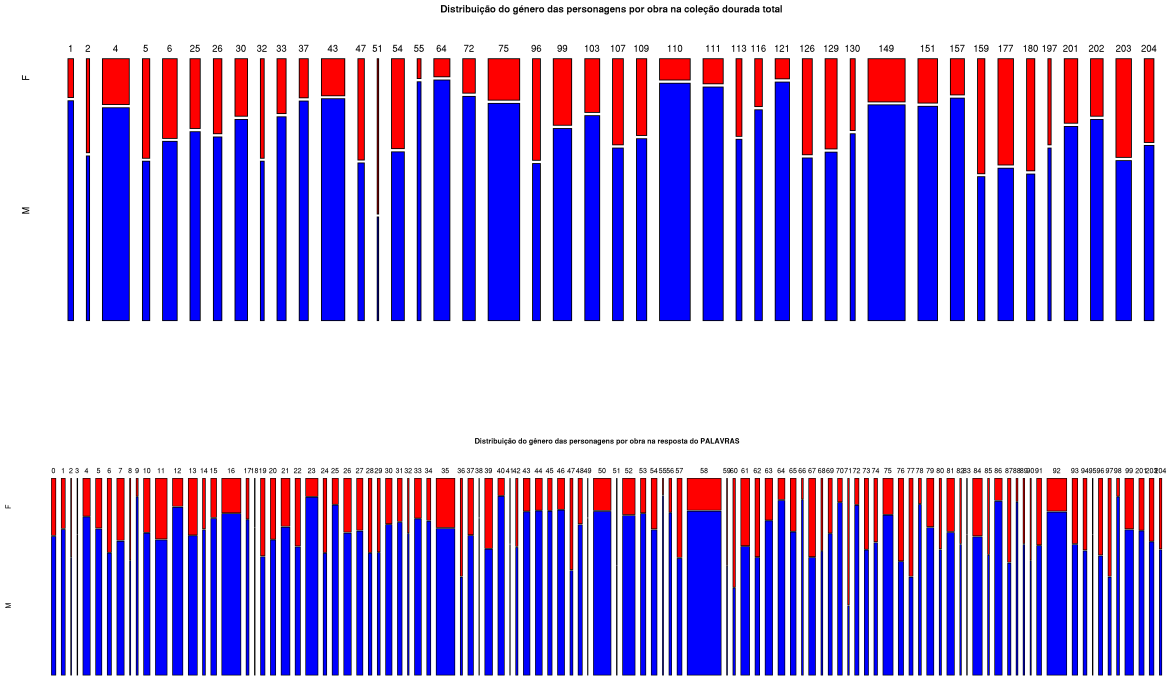
We wanted systems that read the following features of books in Portuguese:

- names (and co-reference)
- gender
- profession/occupation/social status
- family relations with other characters

We provided 200 works, and had done the work "manually" for 40 books (plus 4 examples).

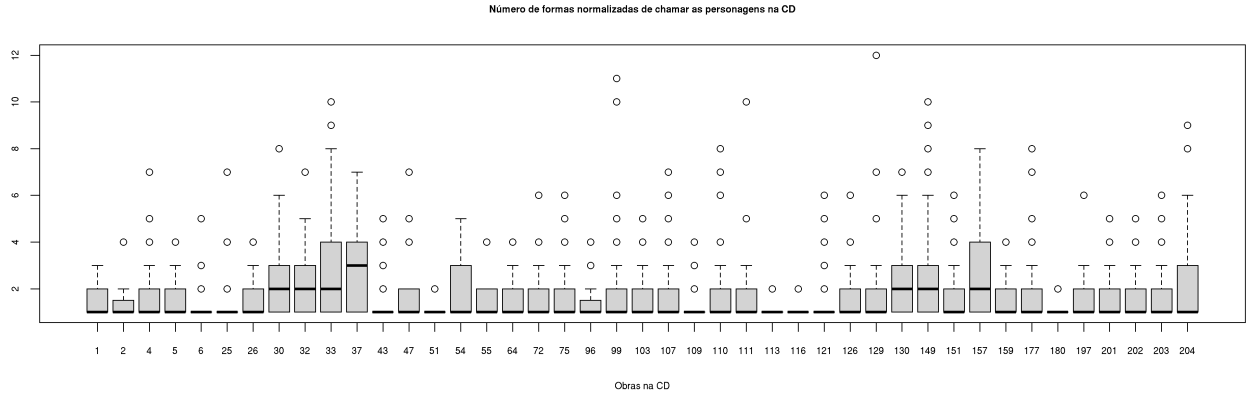
Some results of DIP: character gender

Almost always more masculine than feminine characters in the golden colection (44 works,), and in PALAVRAS's answer (100 works, 4536 male and 1491 female characters):



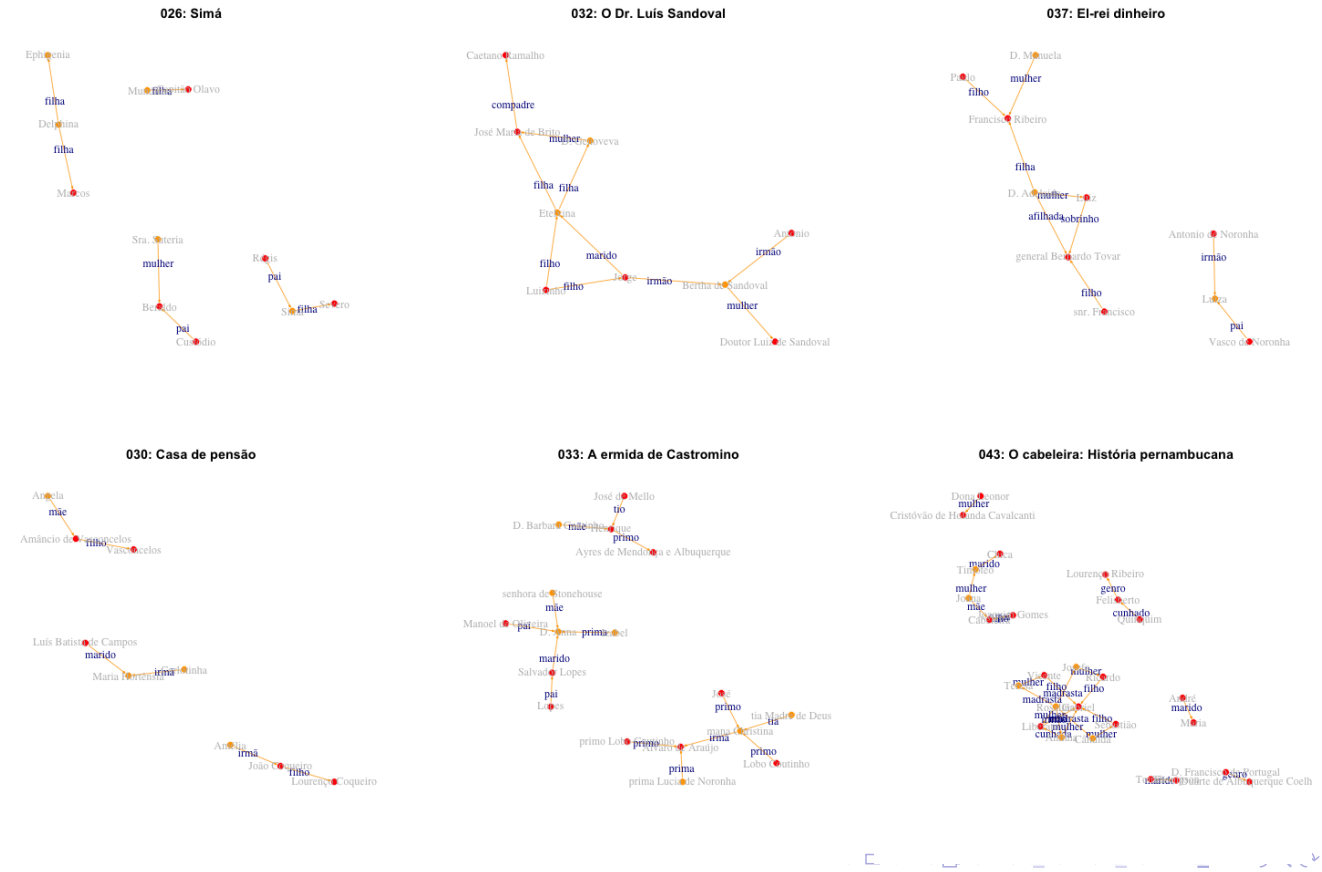
Some results of DIP: names and forms of address

For all novels in the golden collection, there was at least one character referred by more than one form.



Most common female name: *Maria*
Most common male name: *João/Pedro*
80 diminutives: 44 male and 36 female. 36 in text from Portugal, 44 in text from Brasil.

Family relations, Mota (2022)

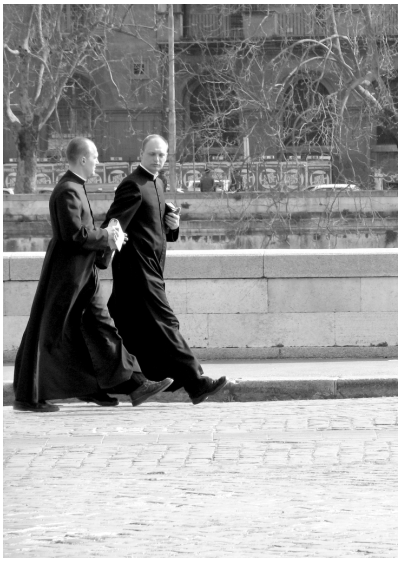


Some results of DIP: understanding the task and the problem(s)

Probably, it was even more important the discussion and clarification of the concepts and the tasks, than the actual results:

- what is a literary character?
- what is a profession/occupation/social status?
- which family relations can/should one identify?

Can we progress from here?



Characterizing people (Freitas & Santos, subm)

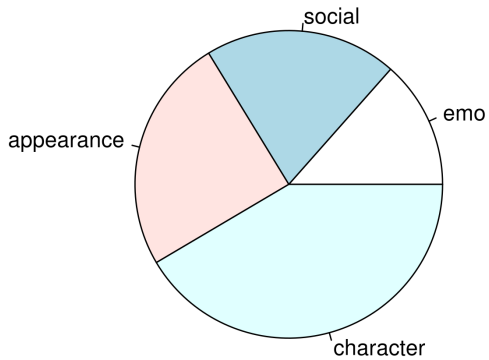
character	*bom* ('good'), *grande* ('great'), *honrado* ('honourable, honest'), *simples* ('simple'), *excelente* ('excellent'), *digno* ('with dignity'), ...
appearance	*velho* ('old'), *novo* ('young'), *antigo* ('old') , *jovem* ('young'), *belo* ('beautiful'), *formoso* ('handsome'), *bonito* ('beautiful'), *lindo* ('beautiful'), *alto* ('tall'), ...
social	*rico* ('rich'), *ilustre* ('illustrious'), *nobre* ('noble'), *casado* ('married'), *célebre* ('famous'), *pobre* ('poor'), *famoso* ('famous'), *poderoso* ('powerful'), ...
emotional	*pobre* ('poor'), *infeliz* ('unhappy'), *feliz* ('happy'), *triste* ('sad'), *alegre* ('joyful'), *amigo* ('friend/friendly'), *apaixonado* ('in love'), *contente* ('contented'), ...

character	1572
social	1472
appearance	669
emotional	502
other	331
total	4253

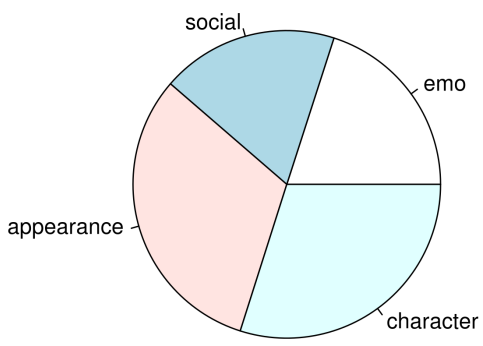
...and looked into how different genders were depicted

	Total	Men	Women
People	482,714	286,721	156,646
Characterised people	42,444	26,301	14,929
Social	7428	4731	2484
Appearance	11,137	6405	4446
Emotion	6891	3777	2976
Character	14,769	9973	4279

Masculine depiction

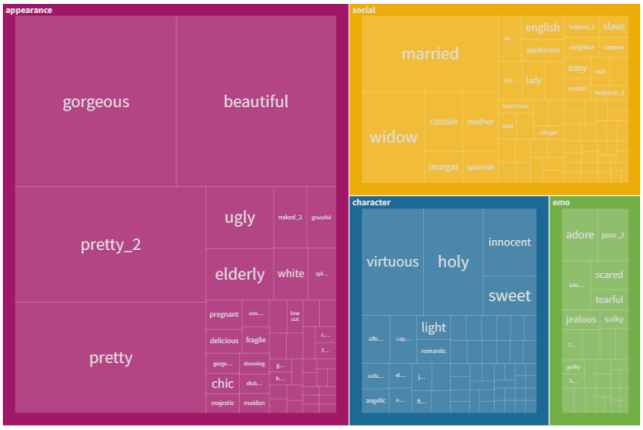


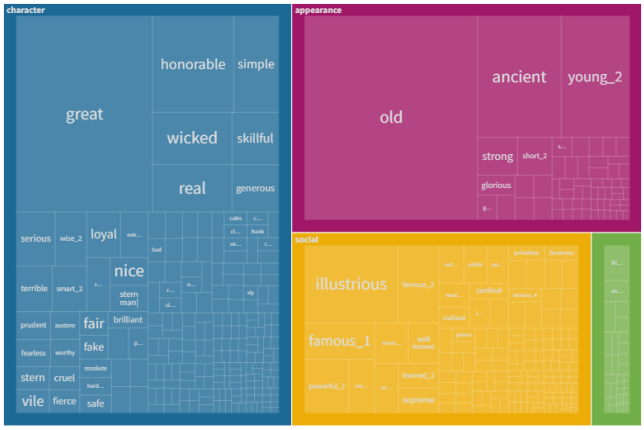
Feminine depiction



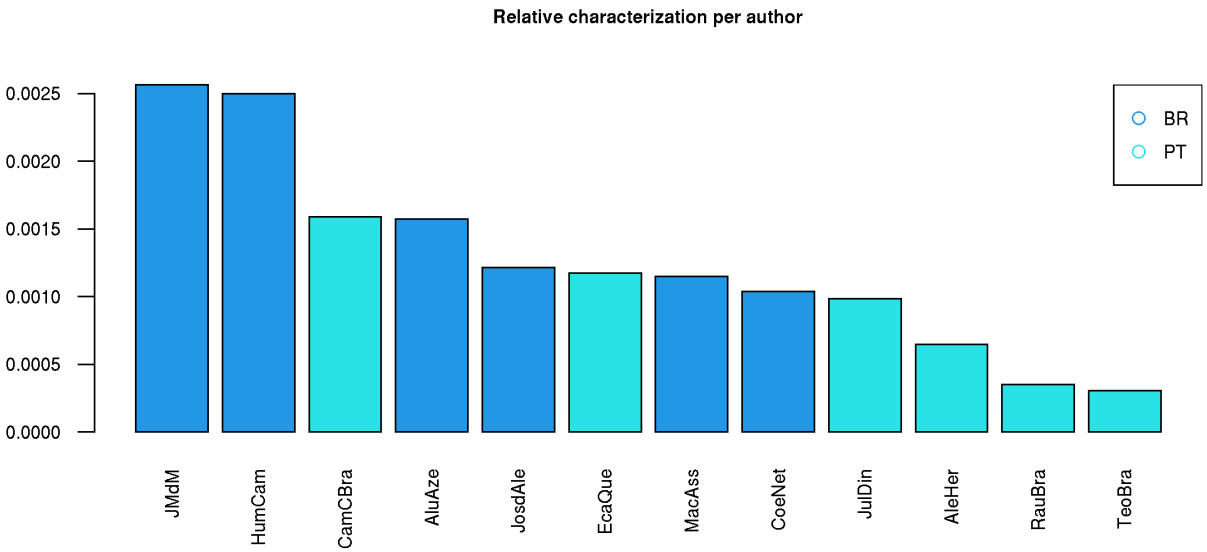
Which characterising words are preferred by gender?

Which words are much more frequently used for feminine vs. masculine descriptions?

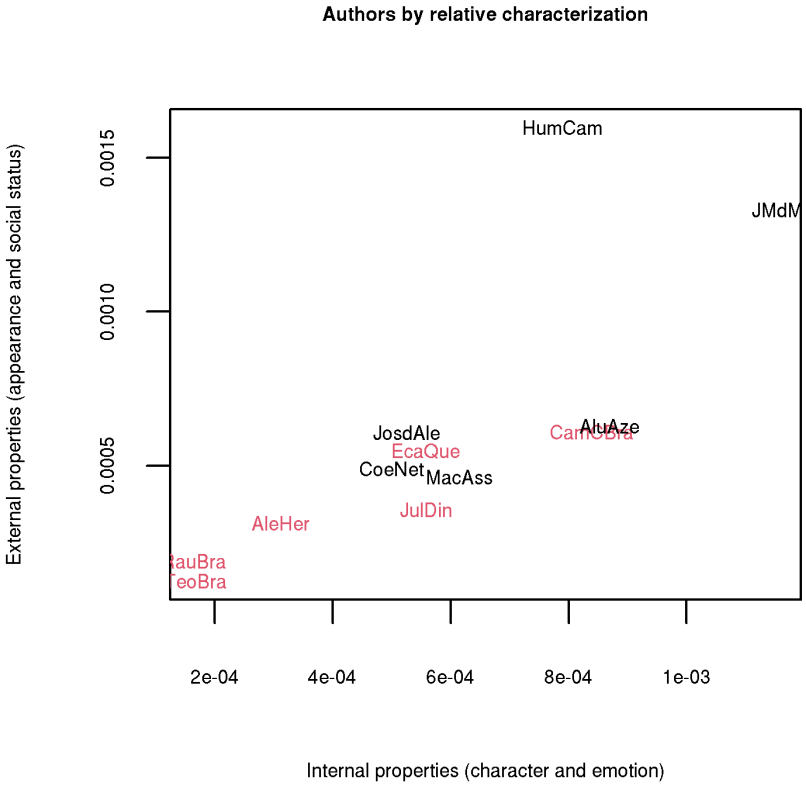




Looking at different authors



Considering external vs. internal characterization



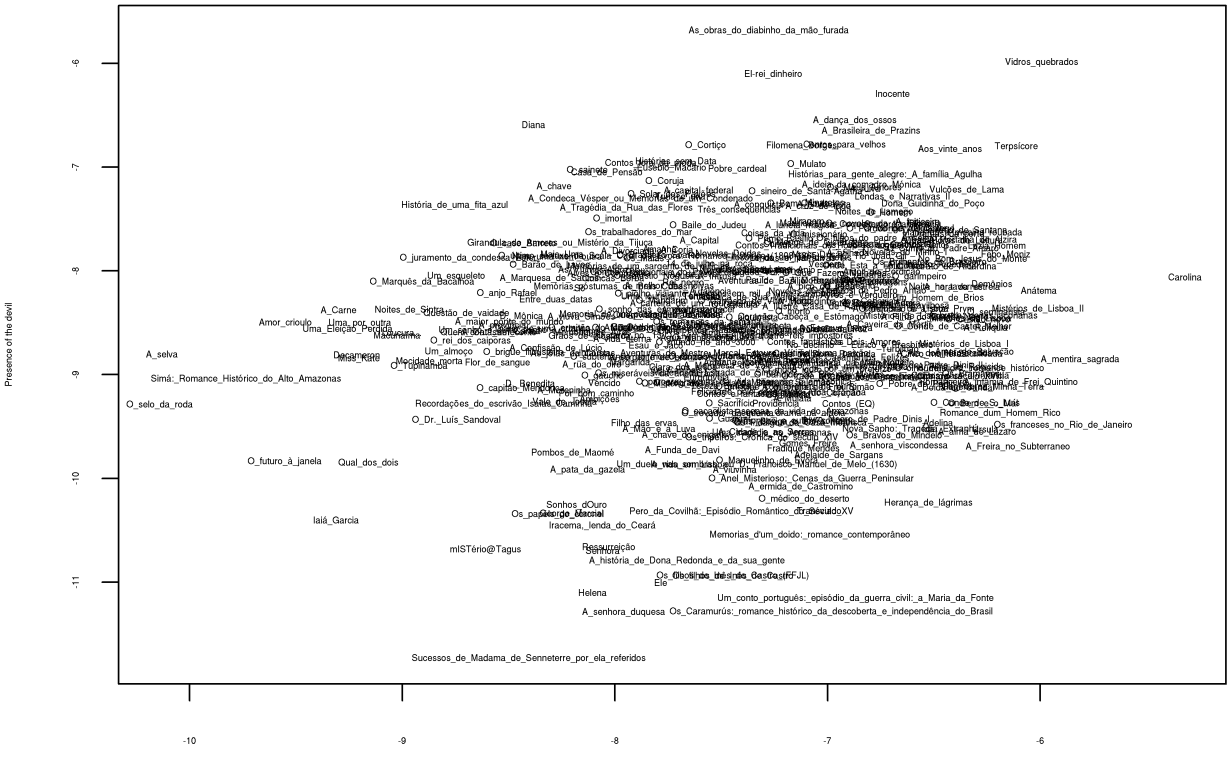
Proper names in general (Santos & Freitas, 2019)

Invoking God and the devil in 259 literary works in Portuguese:

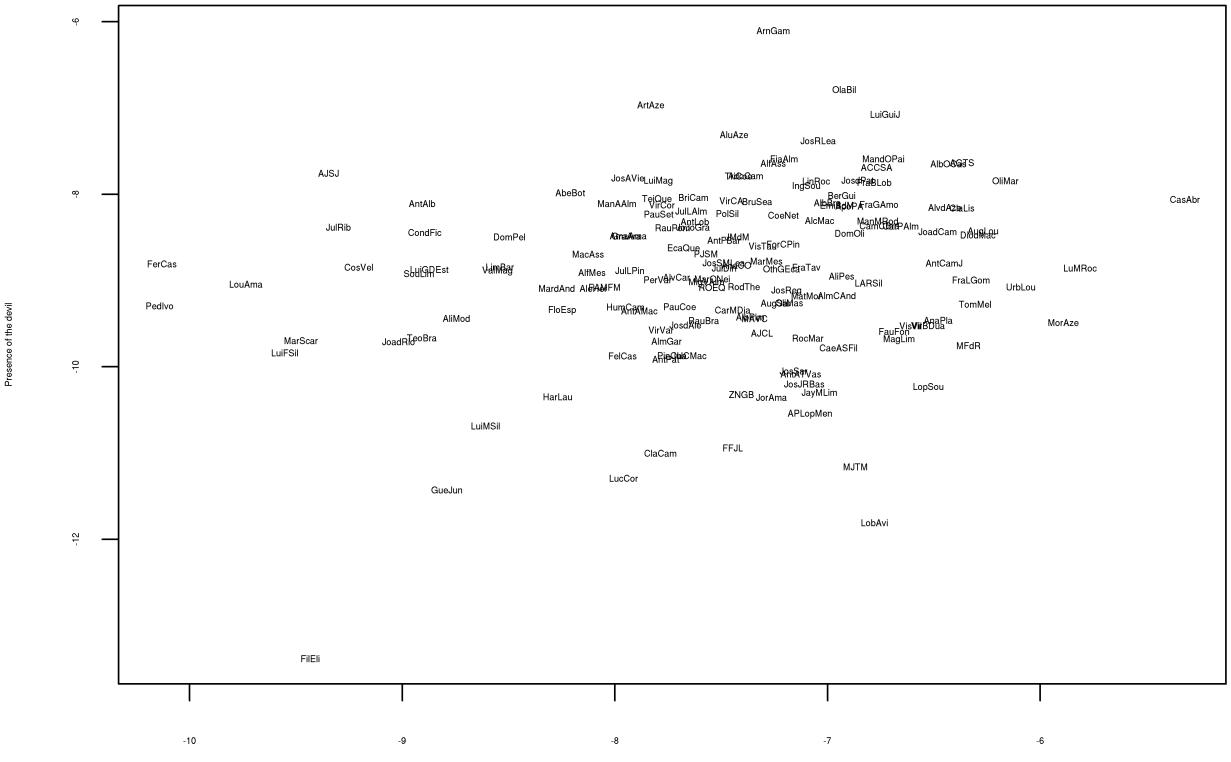
Names for God		names for the devil	
Deus	11.175	Diabo	3.308
Jesus	1.054	Satanás	135
Cristo	677	demo	113
Nosso Senhor	290	Lúcifer	49
Santo Deus	266	Belzebu	14
Nosso Senhor Jesus Cristo	260	Satã	13
Total	13.722	Total	3.632

And with 563 novels, novellas or short stories...

God and the devil, by work (log scale of rel. frequency)



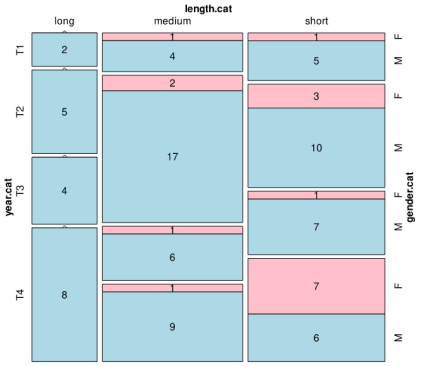
God and the devil, by author (log scale of rel. frequency)

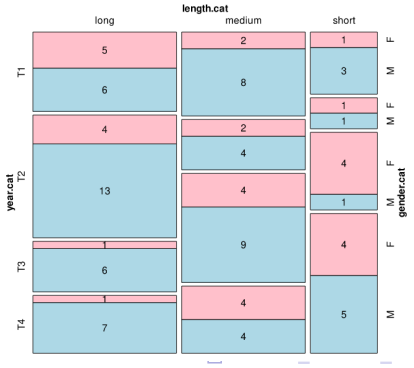


Portuguese and other literatures: the COST action

Distant Reading

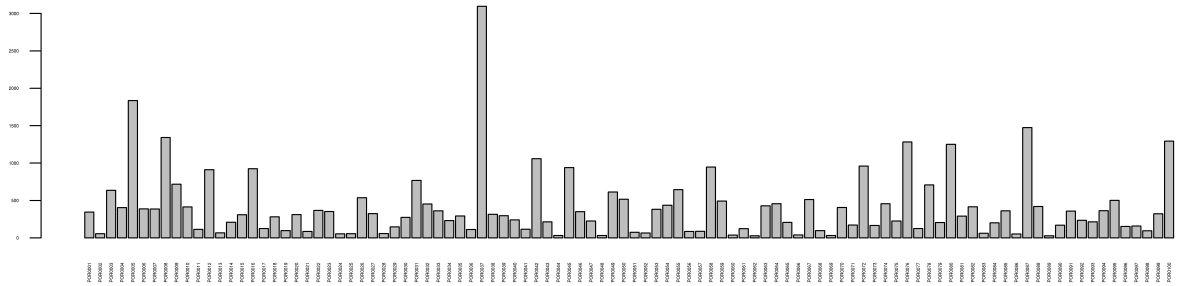
- Creation of several comparable European collections (100 novels each), with novels from 1840 to 1920
- Not an easy task, to ensure comparability and to address all features one wished to compare:
  `https://distantreading.github.io/ELTeC/`
- Especially the operationalisation of "canonicity" was in a way far away from anything we would probably have come up in Portugal
- Different European histories also implied different situations



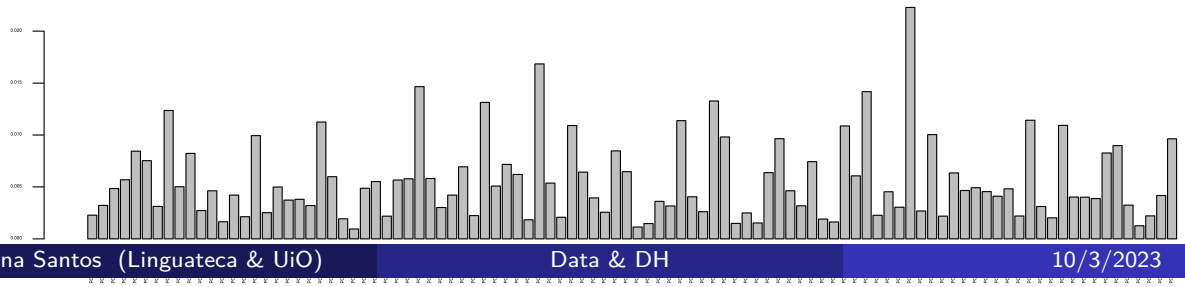


Place names in ELTeC-por (Santos & Bick, 2022)

Absolute: *A Ala dos Namorados* (most) and *Os canibais* (least)

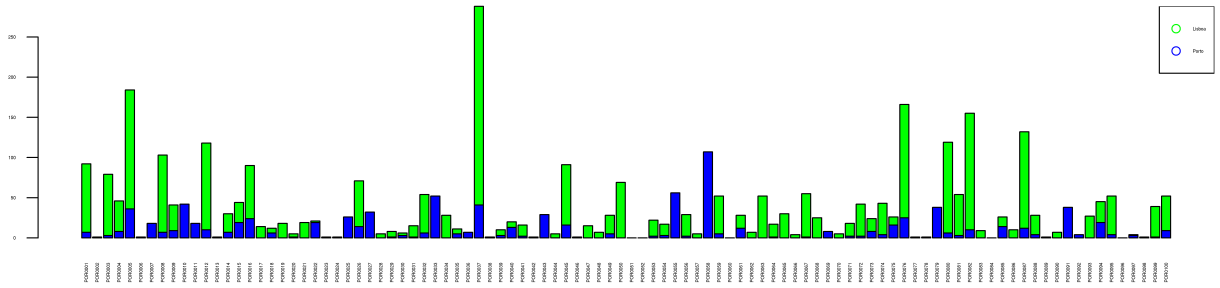


Relative: *No tempo dos franceses* (most) and *A Rosa do Adro* (least)



Presence of Lisbon vs. Porto (absolute)

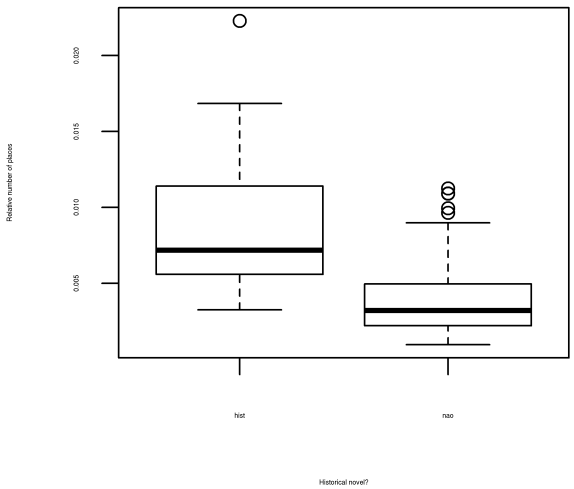
There is often some competition between the capital and the second city, although novelists seem to be well distributed between the two.

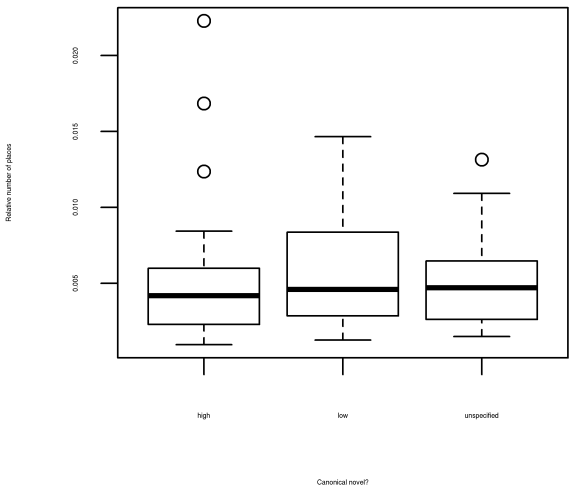


But Lisbon clearly wins in number of mentions.

Do historical novels differ regarding number of places?

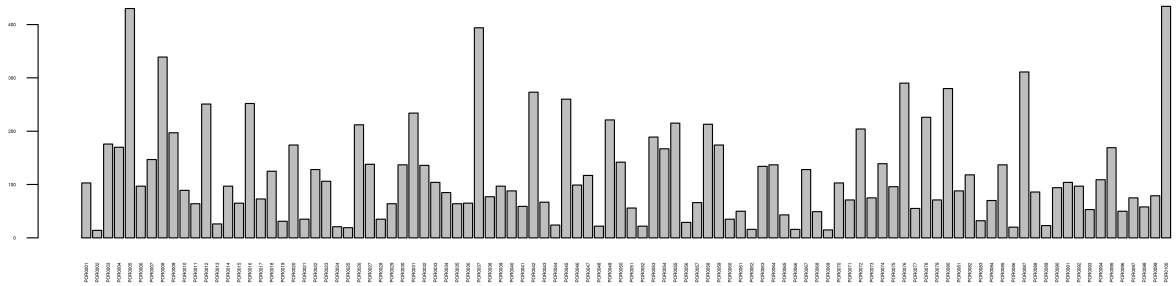
There are 32 historical novels in the ELTeC collection. And there are 26 canonical novels (operationalised by having more than one reprint in the period 1980-2010).

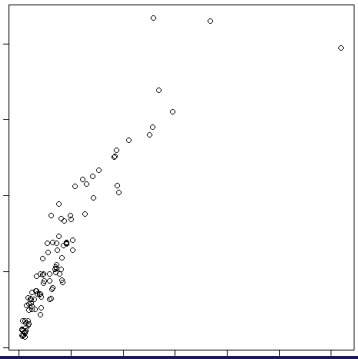




And what about place diversity?

Are there works which are more diverse placewise than others? *A senhora duquesa* (most) and *Sacrificada* (least)

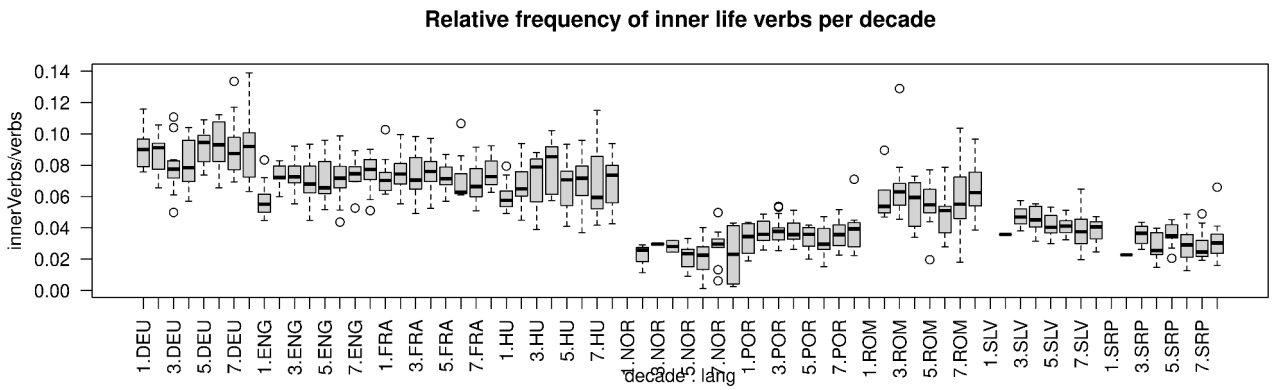




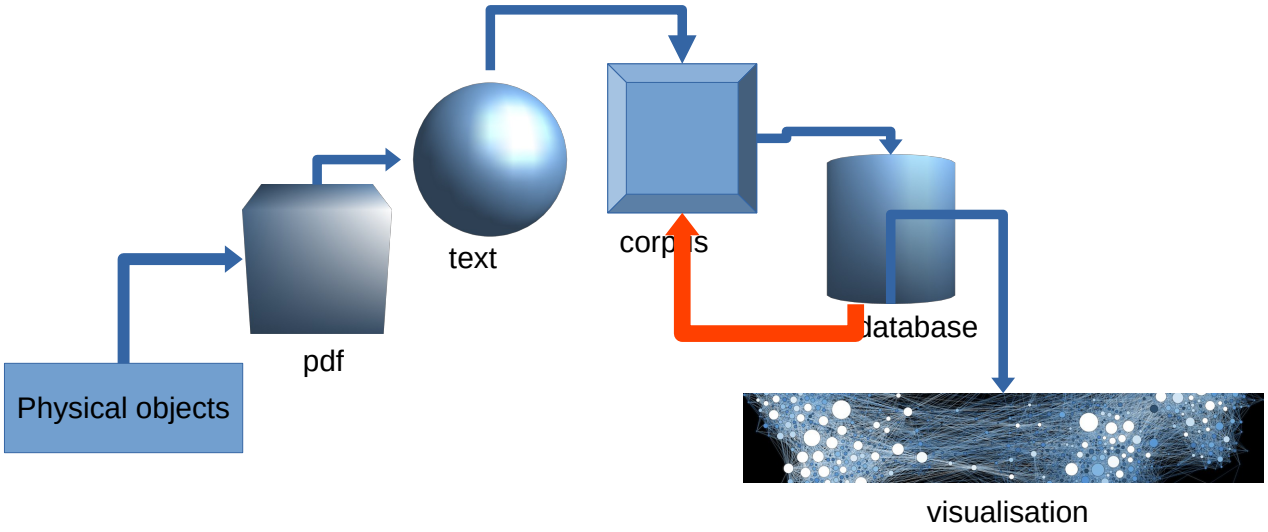
Does the number of distinct places correlate with the number of places? Yes, 0.90

Comparisons with other language collections

- Reference to Christmas, and to Napoleon
- Kinds of named entities
- References to food and drinking (starting now)
- Change in percentage of inner life verbs

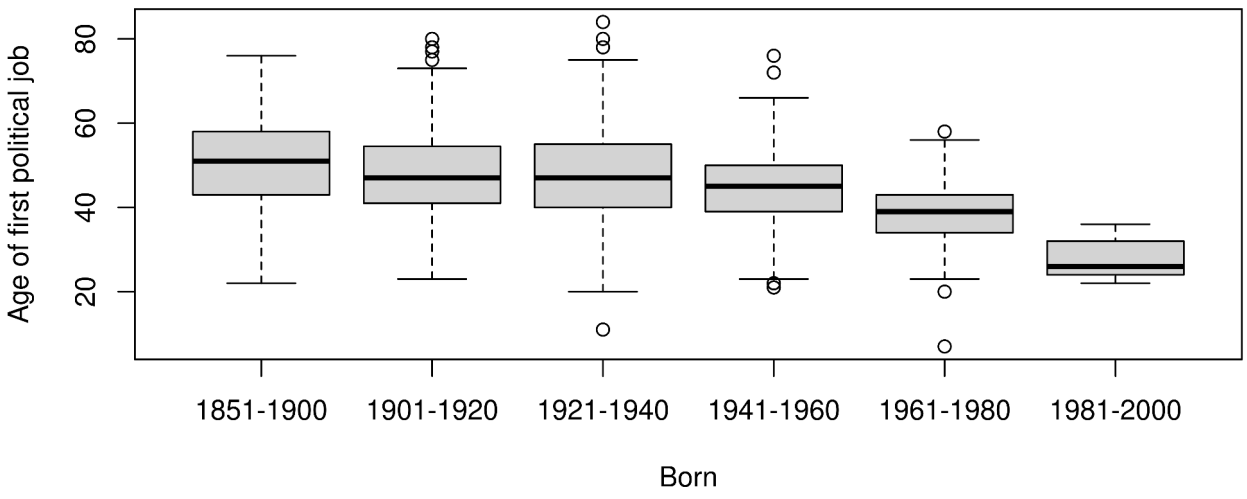


Different phases and objects in DH

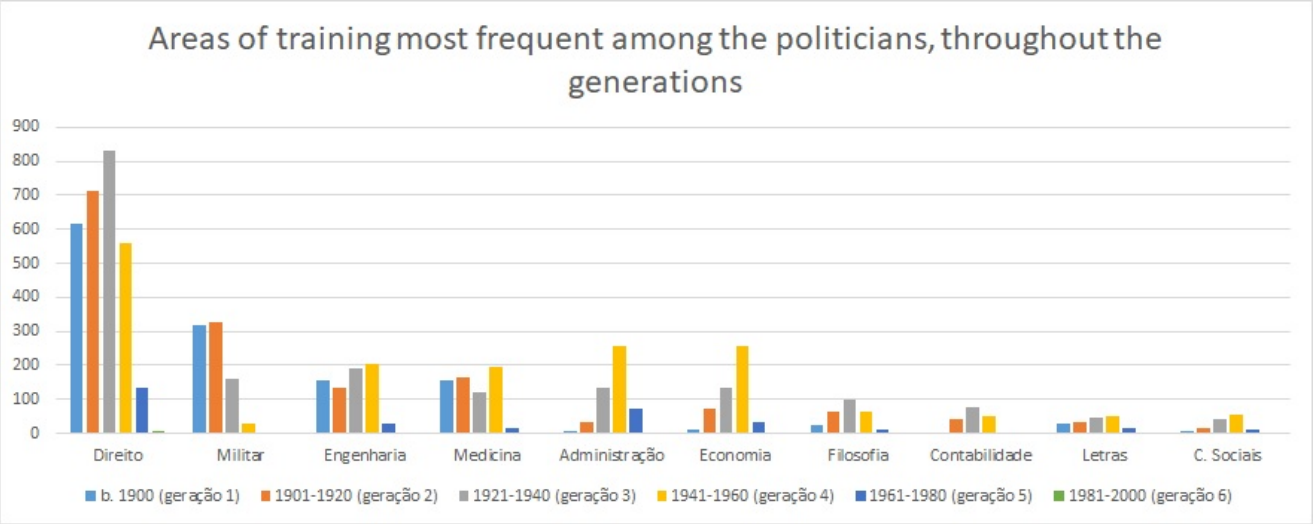


Distant reading Brazilian history (Higuchi et al., 2019)

Using a large encyclopedia on modern Brazilian politics, DHBB, converted to a corpus form to facilitate annotation...

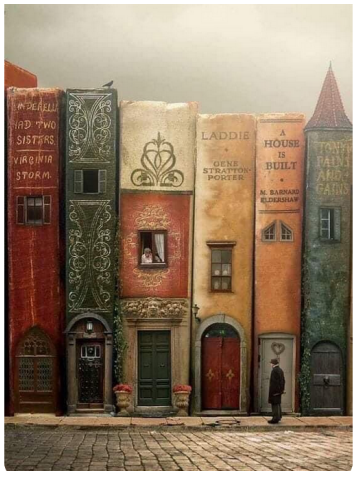


Formal education in DHBB through time (Higuchi et al., 2022)



Concluding remarks

- We need more data for Portuguese – very few literary works are digitized, there are no big digitization initiatives for other kinds of text, copyright is still a big issue
- Still, there is already a lot of research questions that one can try to address, which would be impossible by close reading alone, and these research questions can point to requirements not considered before
- So, it is important to proceed in both directions, so that one can really advance the study of lusophone language, culture and society



https://br.pinterest.com/pin/924012048523164826/

References

- Freitas, Cláudia & Diana Santos. "Human Depiction in Portuguese: Distant reading Brazilian and Portuguese literature". Submitted.
- Higuchi, Suemi, Diana Santos, Cláudia Freitas & Alexandre Rademaker. "Distant reading Brazilian history". In Constanza Navarreta, Manex Agirrezabal & Bente Maegard (eds.) *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference (Copenhagen, Denmark, March 5-8, 2019)*, 2019, pp. 190-200.
- Higuchi, Suemi, Cláudia Freitas & Diana Santos. "Automatic information extraction: a distant reading of the Brazilian Historical-Biographical Dictionary". *PROPOR 2022*, pp. 148-155.
- Mota, Cristina. "Pais, filhos e outras relações no Desafio de Identificação de Personagens (DIP)". *Encontro do DIP*, 21 November 2022.
- Radak, Tamara, Lou Burnard, Pieter Francois, Fotis Jannidis & Diana Santos. "Mapping the Inner Life of Characters in the European Novel between 1840 and 1920". Final action event, 21-22 April 2022.
- Santos, Diana & Cláudia Freitas. "Estudando personagens na literatura lusófona". In *STIL 2019*, pp. 48-52.

References 2

- Santos, Diana, Emanoel Pires, João Marques Lopes, Rebeca Schumacher Fuão & Cláudia Freitas. "Periodização automática: Estudos linguístico-estatísticos de literatura lusófona". *Linguamática* 12 (1), 2020, pp. 80-95.
- Santos et al. "Leitura Distante em Português: resumo do primeiro encontro". *MAT-LIT - Materialidades da Literatura* 8, 1, 2020, pp. 279-298.
- Santos, Diana. "Looking at Named Entities in ELTeC level 2: Training materials". COST Training school, Belgrade, 22-24 March 2022.
- Santos, Diana & Eckhard Bick. "Distant reading places in Portuguese literature". *NorLit2021* (Trondheim, 14-16 June 2022).
- Santos, Diana, Roberto Willrich, Marcia Langfeldt, Ricardo Gaiotto, Cristina Mota, Emanoel Pires & Rebeca Schumacher. "Identifying literary characters in Portuguese: Challenges of an international shared task". *PROPOR 2022*, 413-419.
- Santos, Diana, Cristina Mota, Emanoel Pires, Marcia Langfeldt, Rebeca Schumacher & Roberto Willrich. "Presenting the character identification challenge: setup, resources and results". *Encontro do DIP*, 21 november 2022.
- Santos, Diana & Daniel Alves. "Placing GIS and NLP in literary geography: experiments with literature in Portuguese". *IJHAC: A Journal of Digital Humanities*, 17, 1, 2023, pp. 47-64.