Metadata of the chapter that will be visualized online

Chapter Title	Experiments with Distant Reading … in Portuguese	
Copyright Year	2024	
Copyright Holder	The Author(s), under exclusive license to Springer Nature Switzerland AG	
Corresponding Author	Family Name	**Santos**
	Particle	
	Given Name	**Diana**
	Suffix	
	Division	
	Organization/University	Linguateca and ILOS, University of Oslo
	Address	Oslo, Norway
	Email	d.s.m.santos@ilos.uio.no

Abstract

In this chapter I will discuss a variety of distant reading studies in Portuguese, most of them, but not all, concerning literary works, which have mainly been published in Portuguese.

After an introduction to the concept of distant reading and the macroscope (making things smaller as if increasing the distance, as opposed to microscope), I report on an attempt to predict literary school from linguistic features. In this connection, I deal briefly with the issue of author style by comparing the relative number of nouns compared to verbs. Then, I present a study which analyzes places in Portuguese literature, both real and imaginary places, stressing the complexity of the notion of place in literary text. Next, I address the question of how people are described in lusophone literature by gender, dividing characterizing expressions in four groups: appearance, social status, emotional character, and personality. These are all studies which look at the number of particular features to compare works. The next study takes the character in a novel as unit: I describe a shared task (DIP) whose aim is to characterize literary characters (their gender and profession or social status, as well as how many different names they have), and their family relationships, and how it was the source of many insights on lusophone literature. I will also touch briefly upon some work in progress to measure "inner life" in literature, in the context of the comparison of different European literature through the ELTeC. Finally, targeting a very different kind of text, namely an encyclopedia of Brazilian politicians and politics, I address socio-political studies about Brazil, namely the issue of family relationships among Brazilian politicians.

I conclude the chapter with some remarks on methodology and what seems to be the way forward.

Q1

Q2

AUTHOR QUERIES

Q1 Please provide expansion for
Q2 In

23

Experiments with Distant Reading … in Portuguese

Diana Santos

1 Introduction

This chapter is a short introduction to a group of studies that can be considered distant reading of Portuguese. Most of them, but not all, concern literary works, and have been presented in different venues, often in Portuguese.

By distant reading I mean the use of features to replace the texts, in situations where there are too many texts to be able to process them by close reading. In a way, this is what people have been doing in corpus linguistics all along, but now the emphasis is on units larger than the sentence and on subject matters that transcend or differ from linguistics proper.

It was Moretti [1] who coined the term or at least made it into a hot subject in literature studies. His arguments do not have to concern us here, since many other people followed him in applying "big data"

D. Santos (✉)
Linguateca and ILOS, University of Oslo, Oslo, Norway
e-mail: d.s.m.santos@ilos.uio.no

S. Araújo et al. (eds.), *Digital Humanities Looking at the World*,
https://doi.org/10.1007/978-3-031-48941-9_23

methods to literary text, from cultural analytics to the definition of what is literature, and using increasingly sophisticated statistical methods.

The appearance of topic models [2] was also a key ingredient of distant reading, although it was first proposed for scientific literature or news. It was eagerly adopted by digital historians, so that History is another (humanities) area which is deep into distant reading.

2 Literary School

Our first attempt at looking at a large body of literary texts was literary school prediction, inspired by [3], who distinguished baroque, arcadism, romanticism, and realism in Brazilian literature.

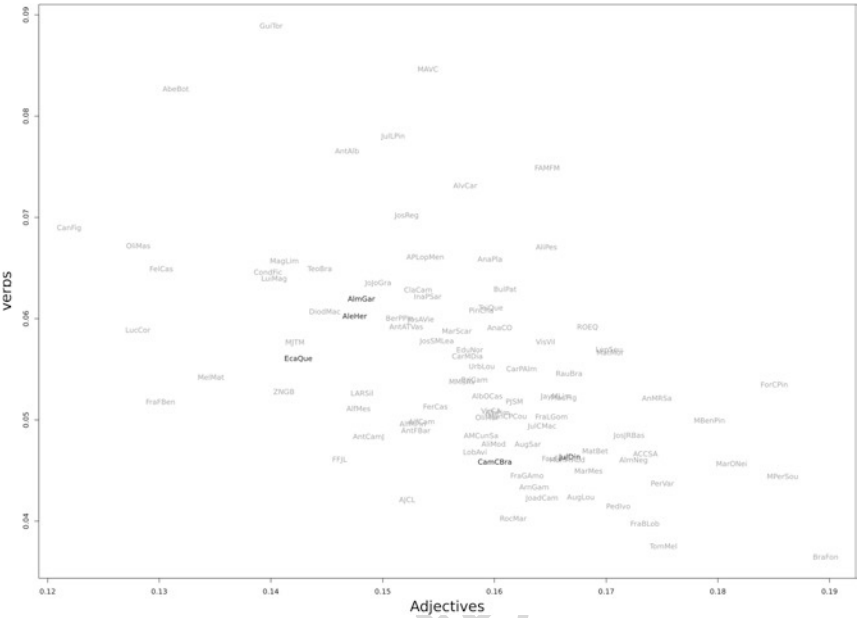
Replacing the texts with a set of linguistic features automatically computed, we wanted to identify the literary schools of works of lusophone literature. In that case we concerned ourselves with decadentism, expressionism, historical, indianism, modernism, naturalism, realism, regionalism, romanticism, and symbolism. We soon found out that literary school was not an easy concept and that there were several disagreements about canonical works. Even worse, there was no algorithm to attribute school to a non-canonical work. This is reported in [4], together with the results obtained.

The style of an author seemed a rather elusive concept, as one can see by simply plotting the density of nouns vs. verbs in Fig. 23.1.

3 Places in Literature

Other studies tried to identify the places in Portuguese literature, both real and imaginary places, see [5, 6]. This made us uncover the complexity of the notion of place in literary text, as well as led to a prototype for automatic map creation, illustrated in Fig. 23.2. We also raised attention to

The advantages of being able to identify places are manyfold, both from a literary perspective, for characterizing genre and authors, and from a historical and geographical perspective, to know more about the



AU4 **Fig. 23.1** Relative use of verbs and adjectives in Portuguese novel authors (in 196 novels)

country (see e.g. [7]), to the most practical application of feeding literary tourism.

4 Depiction of People

How are gender roles created in lusophone literature? To try to answer this question, in [8] we classified human characterizations into four groups, namely appearance, social status, emotional character, and personality, and investigated how these meta-categories relate to masculine and feminine descriptions. In addition, for every characterizing expression which appears more than four times, we identified those which were preferably attributed to masculine and feminine characters. Figure 23.3 shows how differently depicted are the two genders, in a material with mainly fiction before 1920.

Fig. 23.2 Locations in Portugal mentioned in novels by Camilo Castelo Branco. Map automatically computed by the AC/DC interface the many different ways place names were used in text, and revised place annotation according to a set of geographical categories

This sort of task is invaluable for gender studies, in particular to identify changes in the way women have been pictured, as well as uncover possible stereotypes about behavior and social acceptability, to eventually compare with literature in other languages.

Fig. 23.3 Which descriptions are preferred for masculine and feminine characters

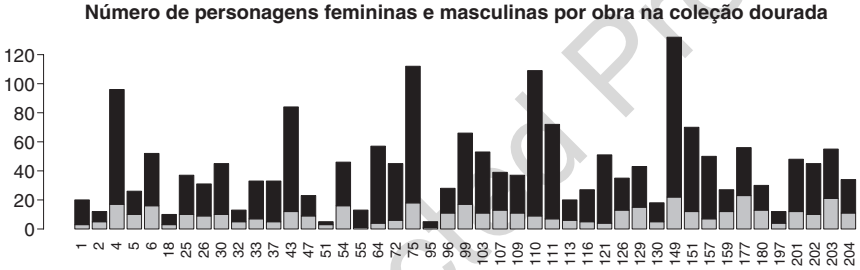


Fig. 23.4 The number of characters per literary work; feminine characters are depicted in grey and masculine in black

AU5

5 Identifying Characters in Literature

Another initiative, with the aim of fostering the development of systems that detected literary characters in lusophone literature was DIP, a shared task to characterize literary characters (their gender and profession or social status, as well as how many different names they have), and their family relationships. See [9] for the announcement, [10] for preliminary results, and [11] for an overview of what was learned with it. In fact, we believe that this organization was the source of many insights on lusophone literature. Let us highlight some findings.

In Fig. 23.4 we see that there are almost always more masculine than feminine characters. Also, feminine characters tended to enter into more family relations than masculine ones, as patent in Table 23.1, adapted

Table 23.1 Number of characters per number of family relations with other characters, out of 252 feminine characters and 823 masculine characters in the golden collection

No. of family rels.	Feminine characters	Masculine characters
None	155	661
1	50	111
2	24	32
3	9	15
4	10	2
5	4	2
All	252	823

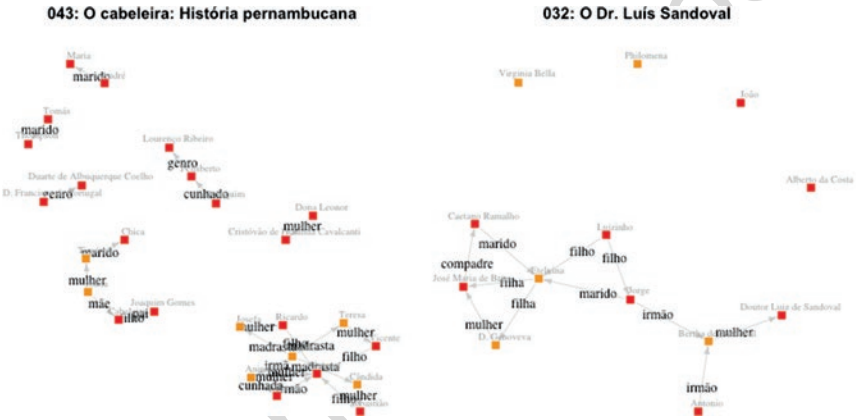


Fig. 23.5 Family relationships in the novels *O Cabeleira: História pernambucana* by Franklin Távora and *O Dr. Luis Sandoval* by Alice Moderno

from [12]. In average, a masculine character has 0.29 family relations, compared to 0.73 of the feminine characters.

In Fig. 23.5 we show different family distributions in different novels (just depicting the characters that have family relationships).

6 Character Networks

Another common task in distant reading is the building of character networks based on co-occurrence in text (not on family relations, as in the previous section). This has also been attempted early on, after human annotation of characters in novels, presented in [13], together with character presence along the book.

Another related concern is what is common across all works, and in Fig. 23.6 we compare, in 259 fiction works in Portuguese, the relative presence of God and the devil—no matter how they are invoked—by author.

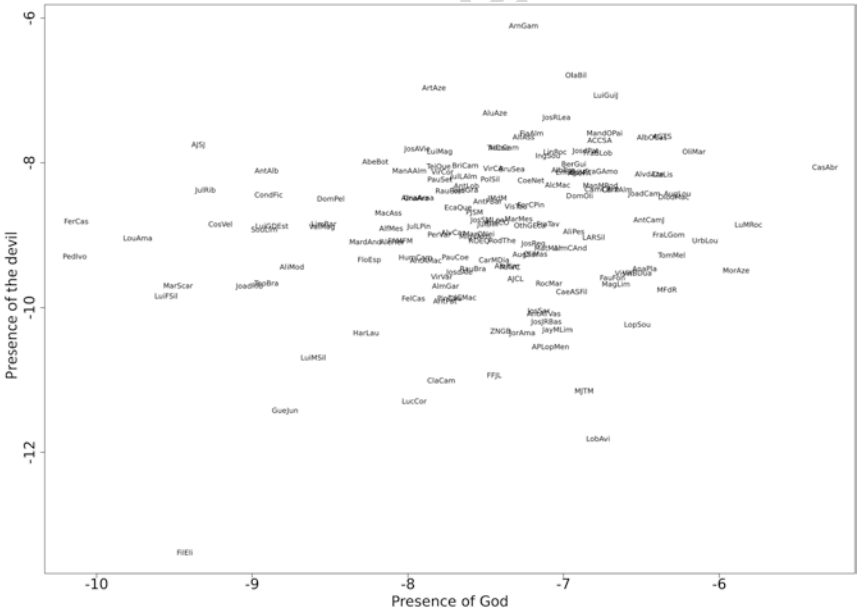


Fig. 23.6 Logarithm of the relative frequency of references to God and the devil per author

7 Inner Life in the European Novel

I should also mention that, in cooperation with experts from other languages and literatures, an attempt to identify changes across European literatures in what concerns inner life description in novels have so far bore no fruits. In [14] we show the results of measuring the proportion of inner life verbs across time in several literatures, using the COST-ELTeC [15].

8 Specific Domains in Brazilian and Portuguese Literatures

One can also compare specific domains in literary text: [16] explores clothing in Brazilian and Portuguese literature, [17] discusses health-related issues, and [18] investigates reference to priests and doctors, among other searches.

9 Brazilian Politics

Finally, let me illustrate the targeting of a very different kind of text, namely an encyclopedia of Brazilian politicians and politics, DHBB (Dicionário Histórico-Biográfico Brasileiro).

In Figs. 23.7 and 23.8 we show some of the subjects that can be asked: when does a politician start her career, and what is the formal education of politicians per generation?

In this work we [19, 20] attempt socio-political studies about Brazil, in particular addressing the issue of family relationships among Brazilian politicians, see [21]. It is interesting to stress that the annotation of family relationships is something that is useful for both studies of fiction and studies of society.

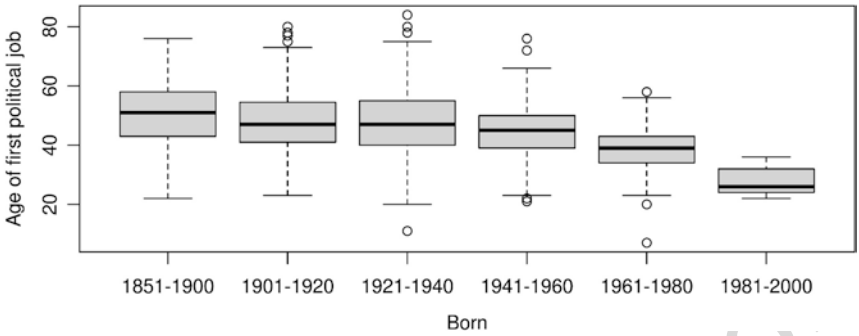


Fig. 23.7 Age of first political job in DHBB per generation

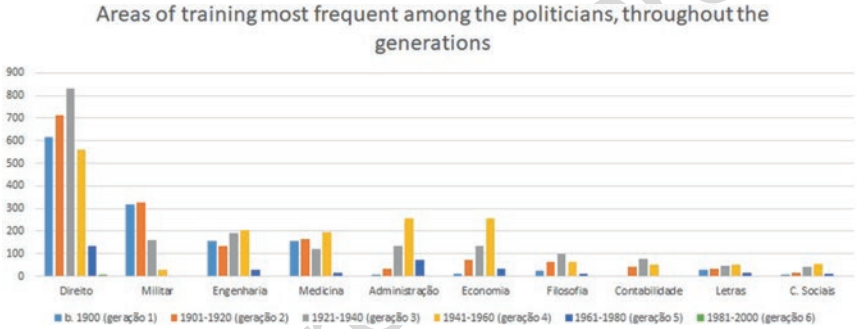


Fig. 23.8 Formal education of Brazilian politicians throughout the twentieth century

this figure will be printed in b/w

10 Concluding Remarks

In a way, most of the examples presented here seem just plain old text mining. However, my proposal is that the results of mining, by being injected back into the corpus annotation, create richer resources the more they are explored and exploited.

I suggest calling this corpus-based digital humanities, illustrated in Fig. 23.9, in which the arrow from database to corpus represents the reannotation of the corpus. One might notice that the studies discussed here have all targeted a small number of books (by today's standards). In that respect, one must acknowledge the blatant lack of digitized literature

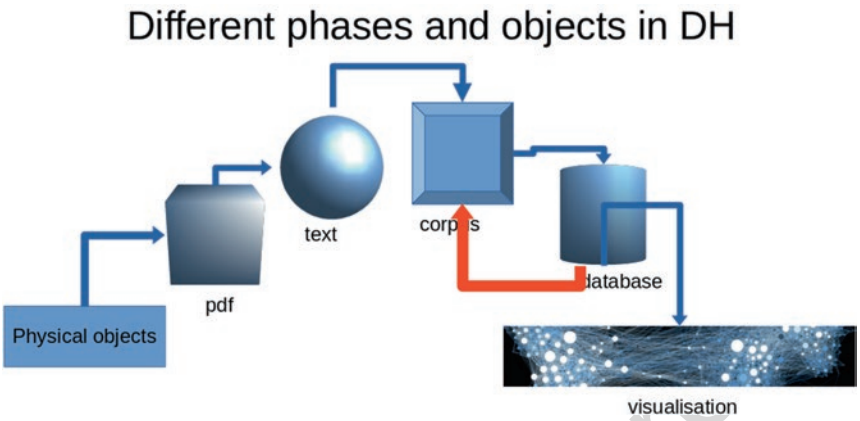


Fig. 23.9 Corpus-based digital humanities: putting back the mined results into the corpus

in Portuguese. We really lag behind languages like Polish or Hungarian, whose countries have digitized thousands of books, or Norwegian, whose national library digitized all books ever published in Norway.

On the other hand, with the existing material several research questions can already be addressed, and hopefully they can point the path for further digitization, making clear requirements for future initiatives—OCR quality being definitely one of them.

I end this short text by suggesting all interested in distant reading in Portuguese to use the Literateca and DHBB corpora for their studies and also of course creating other resources for Portuguese.

Acknowledgments The work described here is joint work (mainly) with my colleagues (in alphabetical order) Alberto Simões, Cláudia Freitas, Cristina Mota, Daniel Alves, Eckhard Bick, Emanoel Pires, João Marques Lopes, Marcia Langfeldt, Rebeca Schumacher, Roberto Willrich, and Suemi Higuchi, to whom I am most grateful. I thank UNINETT Sigma2—the National Infrastructure for High Performance Computing and Data Storage in Norway—for use of their computational resources, as well as Fundação Científica para a Computação Nacional for hosting Linguateca in their servers.

References

1. F. Moretti, *Distant Reading*, Verso, 2013.
2. D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet Allocation, *Journal of Machine Learning Research* 3 (2003) 993–1022.
3. B. Barufaldi, E. F. Santana, J. R. B. B. Filho, J. K. van der Poel, M. M. Júnior, L. V. Batista, Classificação automática de textos por período literário utilizando compressão de dados através do ppm-c, *Linguamática* 2 (2010) 35–44.
4. D. Santos, E. Pires, C. Freitas, R. S. Fuão, J. M. Lopes, Periodização automática: Estudos linguístico-estatísticos de literatura lusófona, *Linguamática* 12 (2020) 81–95.
5. D. Santos, D. Alves, Placing GIS and NLP in literary geography: experiments with literature in Portuguese, *IJHAC: A Journal of Digital Humanities* 17 (2023) 47–64.
6. D. Santos, E. Bick, *Distant reading places in Portuguese literature*, 2022. https://www.linguateca.pt/Diana/download/SantosBickNorLit.pdf.
7. D. Alves, A.I. Queiroz. Studying urban space and literary representations using GIS: Lisbon, Portugal, 1852–2009. *Social Science History* 37, 4 (2013) 457–481.
8. C. Freitas, D. Santos, Gender Depiction in Portuguese: Distant reading Brazilian and Portuguese literature, in: *2nd Annual Conference of Computational Literary Studies*, 2023.
9. D. Santos, R. Willrich, M. Langfeldt, R. G. de Moraes, C. Mota, E. Pires, R. Schumacher, P. S. Pereira, Identifying literary characters in Portuguese: Challenges of an international shared task, in: V. Pinheiro, P. Gamallo, R. Amaro, C. Scarton, F. Batista, D. Silva, C. Magro, H. Pinto (Eds.), *Computational processing of the Portuguese language, 15th International Conference, PROPOR 2022*. Fortaleza, Brazil, March 21–23, 2022 Proceedings, Springer, 2022, pp. 413–419.
10. D. Santos, C. Mota, E. Pires, M. C. Langfeldt, R. S. Fuão, R. Willrich, *Introduction to DIP: goal, setup, resources and results*, 2022. URL: https://www.linguateca.pt/aval_conjunta/dip/apr_encontro/DIPpresentation.pdf.
11. D. Santos, C. Mota, E. Pires, M. Langfeldt, R. S. Fuão, R. Willrich, DIP—Desafio de Identificação de Personagens: objectivo, organização, recursos e resultados, *Linguamática* 15 (2023) 3–30.
12. C. Mota, D. Santos, Pais, filhos e outras relações familiares no DIP, *Linguamática* 15 (2023) 41–53.

13. D. Santos, C. Freitas, Estudando personagens na literatura lusófona, in: *STIL—Symposium in Information and Human Language Technology*, 2019, pp. 48–52.
14. T. Radak, L. Burnard, P. Francois, A. Hilger, F. Jannidis, G. Palkó, R. Patras, M. Preminger, D. Santos, C. Schöch. *Toward a computational history of modernism in European literary history: Mapping the Inner Lives of Characters in the European Novel (1840–1920)*, Open Research Europe, 2023. URL: https://open-research-europe.ec.europa.eu/articles/3-128/v1.
15. C. Schöch, T. Erjavec, R. Patras, D. Santos, Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives, *Modern Languages Open* 1 (2021) 1–19.
16. D. Santos, Explorando o vestuário na literatura em português, *TradTerm* 43 (2021) 622–643.
17. D. Santos, Distant reading health, 2019. URL: https://www.linguateca.pt/Diana/download/DRHealth.pdf.
18. D. Santos, A Gramateca e a Literateca como macroscópios linguísticos, *Domínios de Linguagem* 16 (2022) 1242–1265.
19. S. Higuchi, D. Santos, C. Freitas, A. Rademaker, Distant reading Brazilian politics, in: *Proceedings of 4th Conference of The Association Digital Humanities in the Nordic Countries* (Copenhagen, March 6–8, 2019), 2019, pp. 190–200.
20. S. Higuchi, C. Freitas, D. Santos, Automatic information extraction: a distant reading of the Brazilian Historical-Biographical Dictionary, in: V. Pinheiro, P. Gamallo, R. Amaro, C. Scarton, F. Batista, D. Silva, C. Magro, H. Pinto (Eds.), *Computational processing of the Portuguese language, 15th International Conference, PROPOR 2022*. Fortaleza, Brazil, March 21–23, 2022 Proceedings, 2022, pp. 148–155.
21. D. Santos, S. Higuchi, C. Freitas, Identifying family ties among politicians: Challenges of information extraction evaluation, in: C. Trojahn, M. J. B. Finatto, R. Vieira, V. de Paiva (Eds.), 2nd DH and NLP 2022, *Digital Humanities and Natural Language Processing*, 2022, pp. 69–73.

Author Queries

Chapter No.: 23 0005750838

Queries	Details Required	Author's Response
AU3	In the sentence beginning "In that case we concerned...", note that the spellings have been changed from "expressonism" and "simbolism" to "expressionism" and "symbolism". Please confirm if this is fine.	
AU4	In caption of Fig. 23.1, please check if "in Portuguese novel authors" should be "by Portuguese novel authors".	
AU5	Please confirm if the edits made to the figure caption of Fig. 23.4 are fine.	