Literary similarity among novels in Portuguese

Diana Santos
Linguateca & University of Oslo
Postboks 1003 Blindern, N-0315 Oslo, Norway
d.s.m.santos@ilos.uio.no

Abstract

Through the identification of some features of literary works in Portuguese – linked to their characters, on the one hand; and using syntactic and semantic annotation, on the other – we attempt to study similarity and difference among hundreds of different literary works in Portuguese, using principal components analysis (PCA) to reduce dimensionality. Though a first exploratory study, it already shows some promise. The paper ends explaining the long-term applications we have in mind.

1 Introduction

Can we use data science to identify literary properties, known and unknown? Now that we have access to the data created by the DIP (*Desafio de identificação de personagens*) challenge (Santos et al., 2022, 2023), namely 26 human-revised classifications about particular novels, and ca. 300 automatically annotated by PALAVRAS-DIP (Bick, 2023), called the "extra collection", we can use them as literary motivated features and cluster them.

This is a clear example of distant reading (Moretti, 2013): looking at a large number of books to extract patterns and trends without having to close read them all. And as underlined by Hogan (2011), it is important that new data allow findings that were not considered before, giving rise to new research questions.

The particular motivation for this work is to increase access to literature in Portuguese and make it explorable by the general public and by literary scholars alike. What we present here are the first steps to be embedded in a much larger digital library in the future.

2 Data from DIP

For a sizeable number of works in Portuguese (from Portugal and from Brazil) we have their characters (with all forms used to describe them), together with their gender and profession or social status. In addition, all family relations among characters were also identified.

So, we extracted the following information per work (to make the figures readable, we present the names in parentheses):

- number of masculine characters (`numhom`), number of feminine characters (`nummul`), number of characters (`numpers`)
- number of priests (`padres`), slaves (`escravos`), doctors (`medicos`), kings and queens (`reis`), military professions (`militar`), servants (`criados`)[1]
- number of women with an occupation (`mulprof`)
- number of professions or occupations identified as belonging to a character (`profs`)
- number of characters who are mothers (`maes`), fathers (`pais`), or siblings (`irmaos`)[2]
- number of husbands or wifes (`casais`)

3 Data from AC/DC

But since we had the full text of the works available, we could also compute a set of other features by analysing the text both syntactically and semantically.

More concretetly, all the texts from DIP are also available through the AC/DC project (Santos, 2014), annotated by PALAVRAS (Bick, 2000, 2014), enabling us to obtain several other (possibly) relevant features, such as

- direct speech (no. of –) (direto) and number of speech verbs (`dizer`) (Freitas et al., 2016)

[1] These choices correspond to the most frequent "occupations" discovered in DIP (Santos et al., 2023).

[2] The handling of family relations in DIP included expanding symmetric relations, so if X was mentioned to be e.g. daughter of Y, we would automatically, depending on Y's sex, obtain Y is mother or father of X (Mota and Santos, 2023).

- ratio Perfeito/Imperfeito: namely narrative advancement versus description (`impf.perf`)
- proportion of verbs in the subjunctive mood (`conj`)
- adjective proportion (`adjrel`) (how many adjetives are used in the work compared to the overall number of words) (Santos, 2024)
- average number of words per sentence (`tamfrase`)
- emotion proportion (how many words denote emotions) (`emos`) (Santos et al., 2021)
- how often are some emotions mentioned: love (`amor`), unhappiness (`infeliz`), anger (`raiva`)
- proportion of words from the following semantic domains: clothing (`roupa`), body (`corpo`), health (`saude`), colour (`cor`), ethnicity (`etnicidade`) and family (`familia`)
- density of named places (% of proper nouns as places) (`locais`) (Santos et al., 2020a)
- food and drink mentions (`comida`)

4 Initial analysis

Looking now at how these features locate the different works, using R (R Core Team, 2021) see Figure 1 concerning the 26 works from DIP (we use it for readibility, more figures are available for inspection[3]), it is interesting to note that, for the two authors which have two works in this sample, namely Machado de Assis and Júlio Dinis, their works are quite close when we look at the information coming from DIP. To the lower left can be found short novels written by women (*A vinha*, *Severina*, *A vida por um prejuízo*). On the right lower corner, the three books are historical novels with many characters.

If we look at Figure 2, based on semantic and morpho-syntactic features for those same 26 works, the situation is different: While the works of Machado de Assis are still very close, those of Júlio Dinis get wider apart, apparently because of the difference of importance of the health domain and the named places in the two works, as well as a seemingly higher proportion of direct speech in one of the novels.

We can also investigate the correlation between the two kinds of features, for all books together, through the correlation matrix in Figure 3.

We see that the two kinds of features are quite uncorrelated, which is by itself an interesting result: micro information about the plot and the ambient descriptions is different from high level information as number of characters and their gender, or e.g. how many characters are military. Still, some (tentative) comments can be made: For example, the health domain correlates positively with family relations among the characters. The more such relations exist in the plot, the more health is discussed or mentioned. One may wonder whether the existence of different generations implies that some (old) characters are ill or near death.[4]

Another interesting (negative) correlation is that the more (male) characters, the less emotions are mentioned. This is easy to explain because romantic plots generally have few (sometimes just two) main characters. Plots with many characters are often historical, with many fights and less attention to emotion.

However, mention of anger correlates positively with military characters, kings, and books where most characters have professions – which again sounds like plots with external action and possibly wars and battles.

Conversely, unhappiness correlates negatively with high numbers of characters with professions, and with high numbers of men as characters.[5]

Subjunctive clauses tend to occur with high number of women characters: maybe women are portrayed as uttering more hypothetical sentences or talking more about the future than men?

Clothing is positively correlated with medical doctors and servants as characters. While the second can be associated with dealing with their masters' clothes, it is difficult to understand why novels with medical doctors pay more attention to clothing than others.

Finally, it is interesting to see how colour seems to be unrelated to all other features, which may vindicate the remark in Underwood (2019) that there is no literary theory about colour, urging us to keep with literary-motivated features in distant reading.

To make more clear the weight and relation between the different features, we looked at the loadings on the first three components for just the infor-

[3] https://www.linguateca.pt/documentacao/artigoSemelhanca.html

[4] But this has to be checked.

[5] Obviously, unhappiness and anger are, in a way, opposite emotions, being the first traditionally feminine and the second masculine, so the (almost) perfect inverse behaviour of the two emotions is to expected.

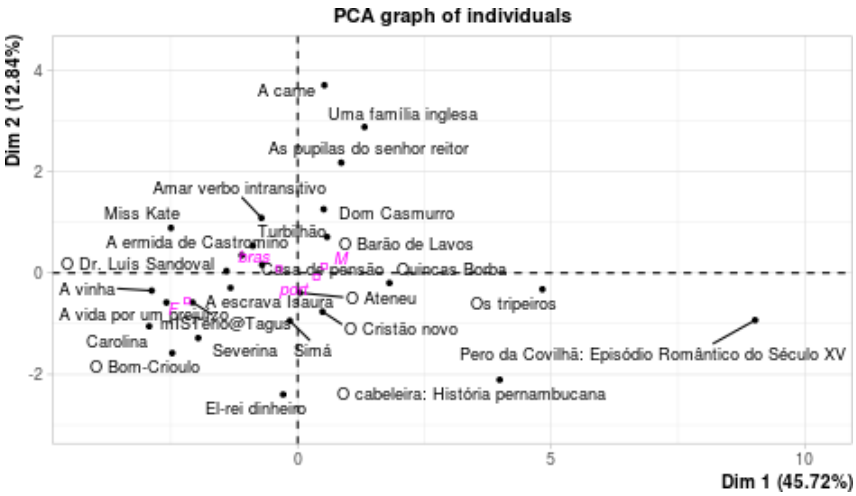

Figure 1: Principal components of 26 works with only the DIP features

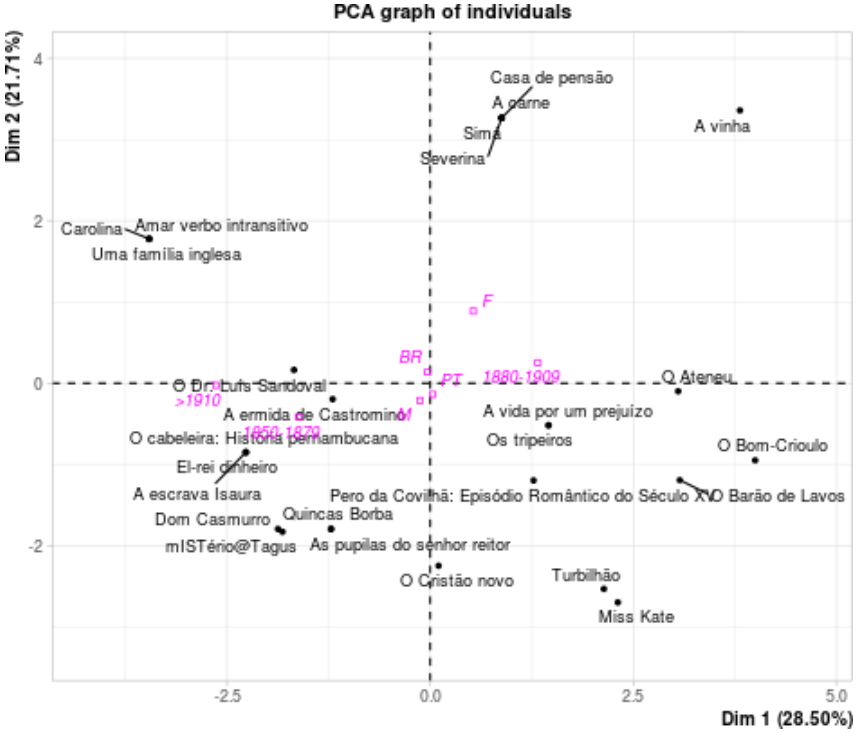

Figure 2: Principal components of 26 works with only the AC/DC features

mation from DIP, just the information from AC/DC, and the two merged, for all works together.

We were able to appreciate that the most discriminative measures are quite varied: the number of characters, the number of professions of the characters, the proportion of saying verbs, the average number of words per sentence, and the proportion of health and ethnicity markers.

This is interesting and should be followed up by more concrete studies on each of these features.

5 Next steps

What we showed was a first clustering based on two different kinds of information about works. We can of course assign hundreds of other low level semantic features to each novel (and will be experimenting with this in the near future). This is work in progress, and we will continue to add information and compute more features to the works and made them public as well. In fact, all data about the novels, in addition to the novels themselves, is publicly available, see URL in footnote 3.

But we would also like to add other kinds of macro level properties, which so far we have no

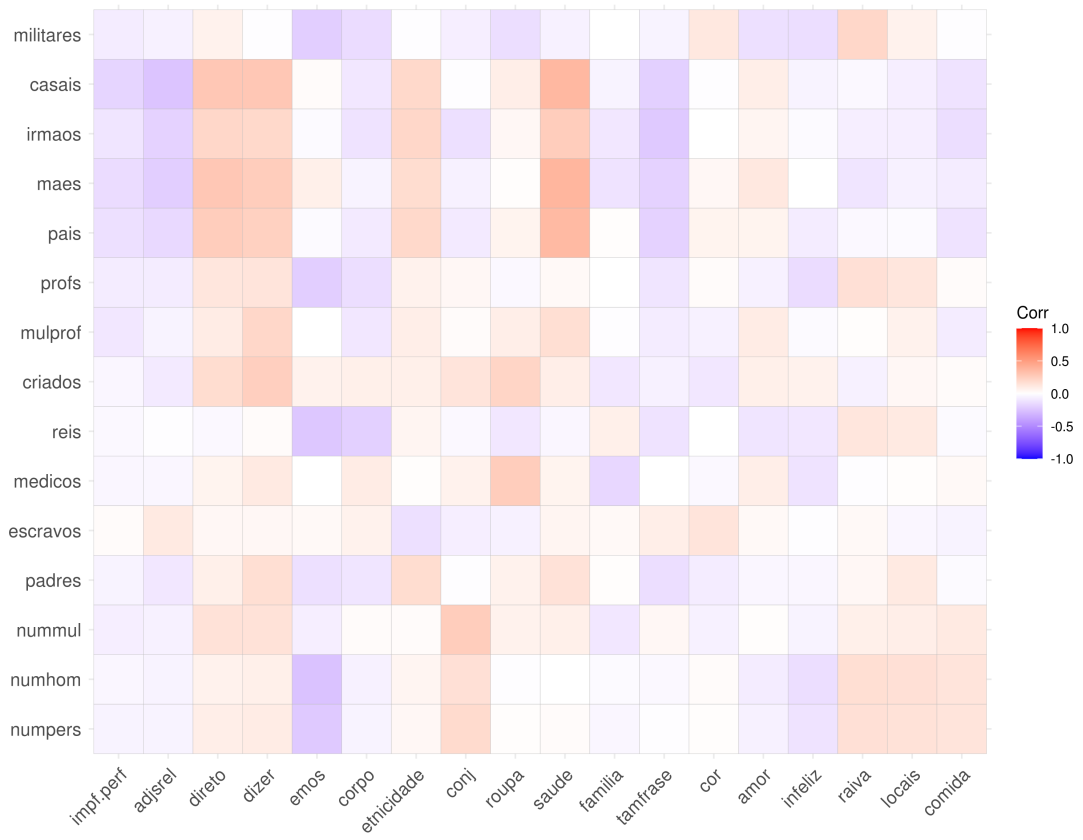

Figure 3: Correlation between the DIP features and the AC/DC features

way to get automatically. Some of them have to do with high level evaluation of a plot, others require more specific knowledge to be gathered. For none of them it is, however, impossible to develop a classifier.

- Does it include an epilogue?
- Kind of ending: happy, tragic, ...
- Environment: field, city, school, sea...
- Does it contain fictive places? Or rather, is it in a "real" place, or in an invented world?
- Social class of the main characters
- Linear time, or flashbacks
- Are children part of the plot?
- Kind of title (names, places, feelings, etc.)

In any case, armed with the knowledge we amass in these exploratory studies, we can develop classifiers to identify

- whether a book is Brazilian or Portuguese
- whether it was written by a man or a woman
- what genre does it belong to (Santos et al., 2020b)
- in which epoch it was written

The two main long-term goals of this work are:

- to develop a "recommender" system pointing to similarities among books in order to suggest new reading experiences in Portuguese, possibly based on a set of questions to the user to identify her preferences.
- to help literary scholars to find points of contact among authors, and get answers to questions about literature history or literary influence, inspired by Archer and Jockers (2016).

We are far from accomplishing either goal, but the work presented here is a required initial step.

Acknowledgements

I am grateful to my colleagues at Linguateca and DIP, withoit whom this paper would npt exist. I acknowledg the Research Computing Services of Sigma, Norway, for cluster facilities, and FCCN - Fundação para a Computação Científica Nacional for the allocation and maintenance of Linguateca's servers

References

Jodie Archer and Matthew L. Jockers. 2016. *The Bestseller Code: Anatomy of the Blockbuster Novel*. St. Martins's Press.

Eckhard Bick. 2000. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Ph.D. thesis, Aarhus University, Aarhus, Denmark.

Eckhard Bick. 2014. PALAVRAS, a Constraint Grammar-based Parsing System for Portuguese. In *Working with Portuguese Corpora*, pages 279–302. Bloomsbury.

Eckhard Bick. 2023. Extraction of Literary Character Information in Portuguese. *Linguamática*, 15(1):31–40.

Cláudia Freitas, Bianca Freitas, and Diana Santos. 2016. QUEMDISSE?: Reported speech in Portuguese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4410–4416.

Patrick Colm Hogan. 2011. *Affective Narratology: The Emotional Structure of Stories*. University of Nebraska Press.

Franco Moretti. 2013. *Distant Reading*. Verso.

Cristina Mota and Diana Santos. 2023. Pais, filhos, e outras relações familiares no DIP. *Linguamática*, 15(1):41–53.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Diana Santos. 2014. Corpora at Linguateca: Vision and Roads Taken. In *Working with Portuguese Corpora*, pages 219–236. Bloomsbury.

Diana Santos. 2024. Experiments with distant reading... in Portuguese. In *Digital Humanities Looking at the World*. Palgrave Macmillan.

Diana Santos, Eckhard Bick, and Marcin Wlodek. 2020a. Avaliando entidades mencionadas na coleção ELTeC-por. *Linguamática*, 12(2):29–49.

Diana Santos, Cristina Mota, Emanoel Pires, Marcia Caetano Langfeldt, Rebeca Schumacher Fuão, and Roberto Willrich. 2023. DIP - Desafio de Identificação de Personagens: objectivo, organização, recursos e resultados. *Linguamática*, 15(1):3–30.

Diana Santos, Emanoel Pires, Cláudia Freitas, Rebeca Schumacher Fuão, and João Marques Lopes. 2020b. Periodização automática: Estudos linguístico-estatísticos de literatura lusófona. *Linguamática*, 12(1):81–95.

Diana Santos, Alberto Simões, and Cristina Mota. 2021. Broad coverage emotion annotation. *Language Resources and Evaluation*, 55(4):857–879.

Diana Santos, Roberto Willrich, Marcia Langfeldt, Ricardo Gaiotto de Moraes, Cristina Mota, Emanoel Pires, Rebeca Schumacher, and Paulo Silva Pereira. 2022. Identifying literary characters in Portuguese: Challenges of an international shared task. In *Computational processing of the Portuguese language, 15th International Conference, PROPOR 2022. Fortaleza, Brazil, March 21-23, 2022 Proceedings*, pages 413–419. Springer.

Ted Underwood. 2019. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press.