Literary similarity of novels in Portuguese

Diana Santos

http://www.linguateca.pt

Distant reading

The goal 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.

Distant reading is a term coined by Moretti (2013) to propose that literary scholars analyse many works to identify trends in literature in general, as opposed to study a small set of canonical works (close reading).

We use automatically obtained features for each work, and have performed several clustering experiments to gauge the interest of the features proposed.

Context

This was possible due to

- the existence of Literateca, a corpus of literary texts in Portugueseo integrated in the AC/DC project, morphosyntactically and semantically anotated
- the previous organization of DIP, *Desafio de identificação de personages* (character identification challenge), in which we had amassed several literary information on several hundred texts.

So, we looked at 238 novels to extract patterns and trends without having to close read them all.

What kind of features?

- from the DIP shared task
  - number of characters, number of male characters
  - number of slaves, number of warriors, number of queens
  - number of parents, couples
  - number of professions, number of characters with multiple professions
- using PALAVRAS annotation and other semantic processing
  - ratio of Imperfeito to Perfeito, relative number of adjectives
  - direct speech, use of subjunctive, number of colours
  - reference to food, health, body or family
  - number of emotions, and specific emotions

Resources

AC/DC (Santos 2014) https://www.linguateca.pt/ACDC/

DIP (Santos et al. 2023) https://www.linguateca.pt/DIP/

Data: https://www.linguateca.pt/documentacao/semelhanca.tsv

Future goals

- 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).

References

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

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

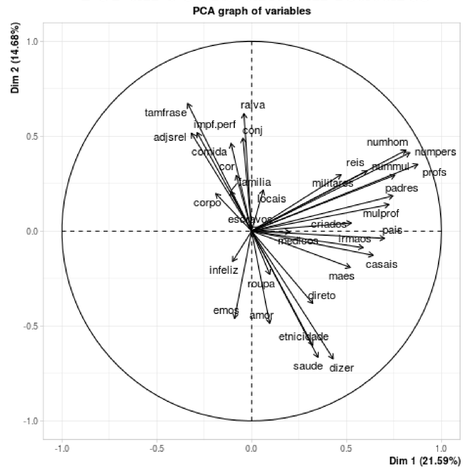
Franco Moretti. 2013. *Distant Reading*. Verso.

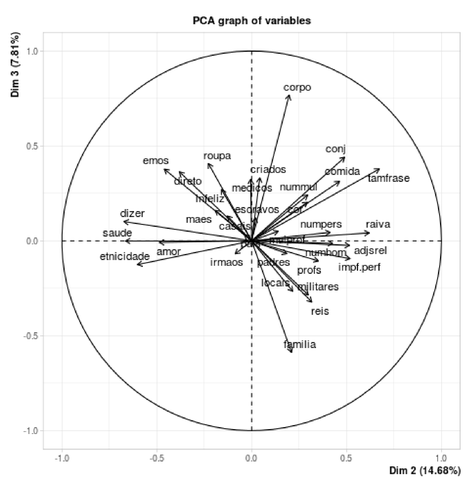
Diana Santos. 2014. Corpora at Linguateca: Vision and Roads Taken. In *Working with Portuguese Corpora*, Bloomsbury, pp. 219–236.

Diana Santos, Cristina Mota, Emanoel Pires, Marcia Caetano Langfeldt, Rebeca Schumacher Fuão & Roberto Willrich. 2023. DIP - Desafio de identificação de personagens: objetivo, organização, recursos e resultados. *Linguamática* 15, 1, pp. 3-30.

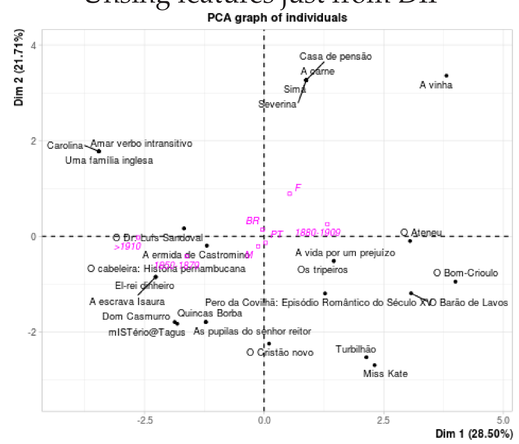
Principal components analysis

For all books and all features

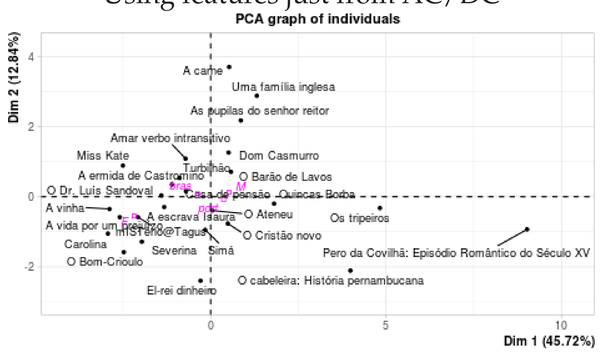




Unsing features just from DIP



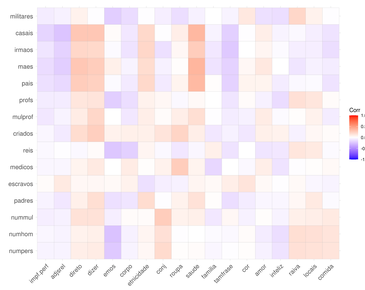
Using features just from AC/DC



In rose, the gender of the autor, the variety, and the period.

Correlation

One of the discoveries was that there was very little correlation between AC/DC features (at the text level) and DIP features (at the plot level), as the following picture shows.



Acknowledgment

Linguateca was financed until 2010 by MCTES, UMIC and FCT, and Linguateca servers are maintained by FCCN. The present work was supported by the University of Oslo. I thank also heartily my colleagues of the organization of DIP and the DIP participants.

UiO : University of Oslo

FCCN Fundação para a Computação Científica Nacional

MCTES Ministério da Ciência, Tecnologia e Ensino Superior

FCT Fundação para a Ciência e a Tecnologia

UMIC Agência para a Sociedade do Conhecimento