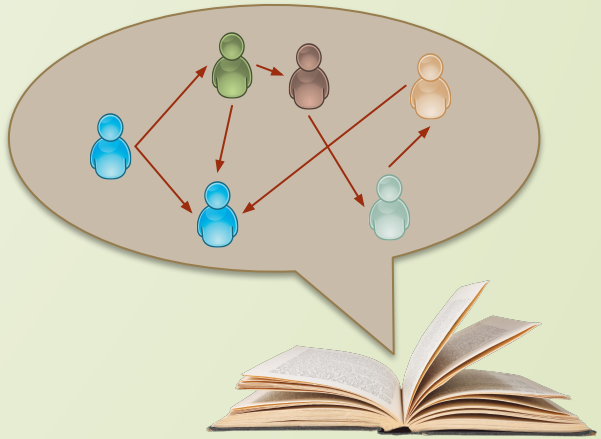
Identifying Literary Characters in Portuguese: Challenges of an International Shared Task

Diana Santos, Roberto Willrich, Marcia Langfeldt, Ricardo Gaiotto de Moraes, Cristina Mota, Emanoel Pires, Rebeca Schumacher Fuão

Motivation

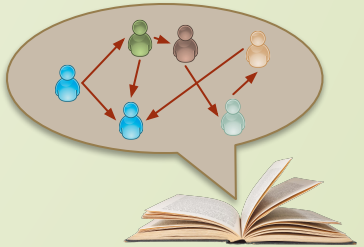
- **DIP Challenge**
  - *Desafio de identificação de Personagens* (Character identification challenge)
  - An evaluation contest to foster the development of systems that, given a literary work in Portuguese, identify and characterize literary characters.
- **Challenge Task**
  - Given a set of literary texts (in text or pdf) the participating systems should provide:
  - the list of characters
  - the different forms they have been referred to in the narrative (as far as proper names are concerned)
  - their gender
  - their profession, occupation or social status
  - family relations between the characters.



What we mean by **character**

- A character is anyone who is mentioned by name in a book and is specific of that book
- Not all people referred in a novel are characters
  - those who do not "advance" the plot, who are not specific of the novel, but represent either the historical scenery or the common culture are not characters

  Examples: Historical people, characters of other well-known stories or religious characters
- DIP scope: we do not try to separate or individualize main characters, secondary characters and so on



DIP Results

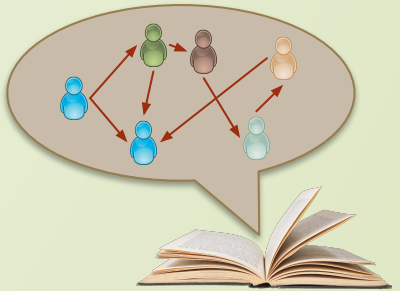
- **Results for Dom Casmurro (incomplete)**

Character	Sex	Profession/Occupation/Social Position
Pedro de Albuquerque Santiago	M	congressperson
Bento | Padre Bentinho | Sr. Bentinho | Doutor Santiago | Dom Casmurro	M	lawyer
Capitu | Capitolina	F	
D. Glória | Dona Glória | D. Maria da Glória Fernandes Santiago | Prima Glória | mana Glória	F	

Character	Relation	Character
Bento	husband	Capitu
Capitu	mother	Ezequiel A. Santiago
D. Glória	widow	Pedro de Albuquerque Santiago

Why are we doing this?

- **There are many digitized books (in public domain)**
  - One can start doing distant reading of a whole literature body/corpus (or at least a sizeable part of it)
  - Read every book to knowledge elicitation is a hard task
- **If we have computerized helpers**
  - We can obtain features from every book
  - Character identification is the first step for distant reading



DIP: Expected contributions

- **Contribute to get high quality systems for character identification in Brazilian and Portuguese literature**
- **Research questions:**
  - what kind of social occupations were dealt with in?
  - how often are slaves mentioned by name?
  - has the percentage of women characters changed?
  - how closely / family related are characters?
  - do number of characters correlate with literary school, or canonicity?
  - how often are kings characters?
- **Other questions**
  - are animals referred by name?
  - are plots urban or rural?
  - are priests or friars or doctors very common?
  - are people from other countries (and with other names) prominent?

Example of challenge

- **From analysis of two works: *As Pupilas do Senhor Reitor* and *Dom Casmurro***
  - There is a clear use of diminutives, but with different purpose
    - Pupilas: they simply indicate a tender way to talk to a particular character, see *Clara*, *Clarinha*, *Clarita*, *Clarita do Meadas*
    - Dom Casmurro: they constitute the form of naming children, and are used to distinguish children from parents, see *Capitu*, *Capituzinha*
- **Challenge:**
  - How to distinguish between different kinds of diminutives
  - Maybe this can show two different traditions/cultures. (Of course, we are NOT saying that this is the case based only on two works)
  - DIP challenge is applied to a large number of books in Portuguese -- may help elucidate this question

DIP Organization

- **Provided data:**
  - DIP organization will provide 200 books in digital form (100 in text and 100 in pdf)
  - Samples of CSV files to be submitted (characters.csv, relationships.csv)
- **Challenge task:**
  - Participants have to deal with, in 48 hours, and return to us two csv files on the 200 files, through EasyChair

DIP Organization

- **Evaluation**
  - Based on a golden collection: 40 literary works
  - For each book, a score is given as the average of the 5 measures:
    - how many different forms of character naming were found?
    - how well they are associated to the same character
    - was the gender of the character correct?
    - were the profession(s), occupation(s) and/or social status correctly found?
    - were the relations between characters correct?
  - The global score is the sum over all 40 books.
- **By the end of DIP**
  - we will make available the golden collection, as a resource for further development.

Final Considerations

- DIP can be the beginning of computational literary studies that can later on find more complicated features
  - such as parts of the plot, (other) relations (such as professional) between characters, or places.
- We hope we raised your interest in DIP
  - The challenge will take place in September 2022
  - we have just finished a dry run ("ensaio") where several people tried their hand on two further novels, manually.
- See http://www.linguateca.pt/DIP for further information.

Motivation

- **For Literary studies**
  - Character identification is as a natural first step of distant reading, given the importance of characters for literary studies
  - Characters sustains the plot and moves it in a particular direction, and may also organize the discourse.
  - Characters are ideological products of authors created in an attempt to revive or project experiences they have
  - Character identification from thousands of works enables to read the (character) landscape by epoch, literary genre, and/or author, expanding our base with many works outside the literary canon
  - The form of the names themselves is relevant, not only because address forms reflect different social status, but because some epithets have relevant interpretations.
- **For Computer Science**
  - A hot research topic: information extraction task
  - Populate a Knowledge base with characters, their attributes, and relations to other characters form literary works in Portuguese.