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

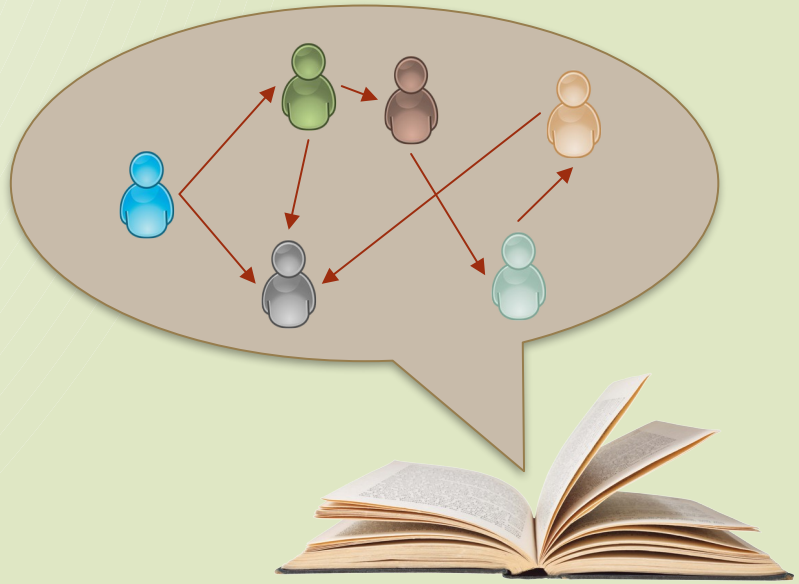
DIP Challenge

*Desafio de Identificação de Personagens (*Character identification challenge)

An evaluation contest to foster the development of systems which identify and characterize literary characters in literary works in Portuguese.

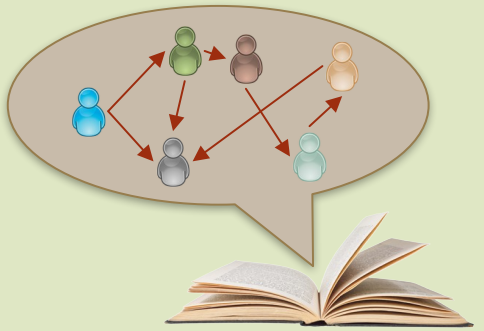
Given a set of literary texts (in text or pdf) the participating systems should provide:

- 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, or advances the plot in the book
- Not all people referred in a novel are characters
  - Historical people, famous people that belong to the culture
  - Literary characters from other works
  - Religious characters
- No attempt to distinguish main characters, secondary characters and so on



DIP: Motivation

- Get high quality systems for character identification in Brazilian and Portuguese literature
- Examples of research questions after these systems exist:
  - what kind of social occupations were dealt with?
  - how often are slaves mentioned by name?
  - has the percentage of women characters changed over time?
  - how closely / family related are characters?
  - does the 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, friars or doctors common?
  - are people from other countries (and with other names) prominent?

Example of a specific question

- From analysis of the two example works: *As Pupilas do Senhor Reitor* and *Dom Casmurro*
  - There is a clear use of diminutives, but with different purpose
    - In *Pupilas*: they simply indicate a tender way to talk to a particular character, see *Clara*, *Clarinha*, *Clarita*, *Clarita do Meadas*
    - In *Dom Casmurro*: they constitute the form of naming children, and are used to distinguish children from parents, see *Capitu*, *Capituzinha*
- After the challenge:
  - Are diminutives conspicuously present in character names? Male and female alike?
  - Can this show two different traditions/cultures of diminutives? (Of course, we are NOT saying that this is the case based only on two works)
  - May other functioons of diminutives emerge?

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.

Concluding: highlights

- For literary studies
  - Character identification is as a natural first step of distant reading, given the importance of characters for literary studies
  - 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 canon
- For computational linguistics
  - The form of the names themselves is relevant, not only because address forms reflect different social status, but because some ways of addressing can 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.