Identifying literary characters in Portuguese
Challenges of an international shared task

Diana Santos[1⋆][0000−0002−3108−7706], Roberto Willrich[2][0000−0001−5067−0406], Marcia Langfeldt[1][0000−0002−1600−9769], Ricardo Gaiotto de Moraes[3][0000−0003−3595−0033], Cristina Mota[4][0000−0001−8127−8503], Emanoel Pires[5][0000−0001−7377−8540], Rebeca Schumacher[1][0000−0002−7658−7704], and Paulo Silva Pereira[6][0000−0001−9995−4063]

[1] Linguateca & University of Oslo, Norway
{d.s.m.santos,r.s.e.fuao}@ilos.uio.no,marcia.langfeldt@gmail.com
[2] INE-UFSC, Brazil
roberto.willrich@ufsc.br
[3] DLLV-UFSC, Brazil
rgaiotto@gmail.com
[4] Linguateca & INESC-ID, Portugal
cmota21@gmail.com
[5] UEMA-UFPI, Brazil
emanoel.uema@gmail.com
[6] Universidade de Coimbra, Faculdade de Letras, Portugal
psilvapereira@sapo.pt

Abstract. We introduce the problem of identifying characters in literary text, and mention some specific issues that are special for Portuguese, in the context of presenting DIP, a shared task to foster work in the area and produce resources for computational literature studies in Portuguese. We describe how the task is organized, the resources that will be created, and how we plan to evaluate the results produced by the participant systems.

Keywords: Lusophone literature · Distant reading · Digital humanities

1 Identifying characters in literary text

We describe here DIP: *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, see hhtps://www.linguateca.pt/DIP/.

We see this as a natural first step of distant reading in Portuguese, given the importance of characters for literary studies. In DIP we will deal with novels (written in the last 250 years), digitized as pdfs, or already in text form.

⋆ Author to whom correspondence should be addressed

1.1 Brief motivation from literary studies

A literary character is important in fiction, as it sustains the plot and moves it in a particular direction, and may also organize the discourse (if the narrator is also a character). Since they are created by the author in an attempt to revive or project experiences, they are ideological products. And characters are then built by the reader depending on her beliefs and context, so their reception can drastically change depending on the audience.

If we can get information on characters from thousands of works, we may be able to read the (character) landscape by epoch, literary genre, and/or author, expanding our base with many works outside the literary canon, which may provide interesting opportunities for postcolonial, gender and queer studies. Specifically in a Brazilian context, the presence (or absence) of slaves as characters in the literature is a most relevant concern.

Finally, the form of the names themselves is relevant, not only because address forms reflect different social status, but because some epithets have relevant interpretations, as the case of *Capitu* and *Bentinho* in Dom Casmurro, see [5].

1.2 Motivation from computer science studies

From a computer science perspective, one can see the problem as a standard information extraction task, that from literary works must populate a knowledge base with characters, their attributes, and relations to other characters, as illustrated in Figure 1.

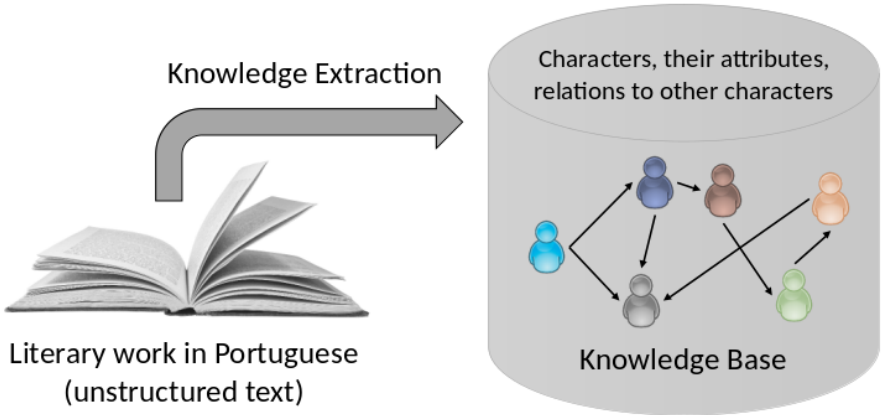

Fig. 1. The task of DIP in a nutshell

In order to perform this extraction task, one may have to use various NLP techniques, which make the challenge more related to (somehow) understanding Portuguese text. Figure 2 presents the main steps in DIP.

The first is named entity recognition of the person names present in the narrative. But it should be noted that DIP is not interested in all person names,

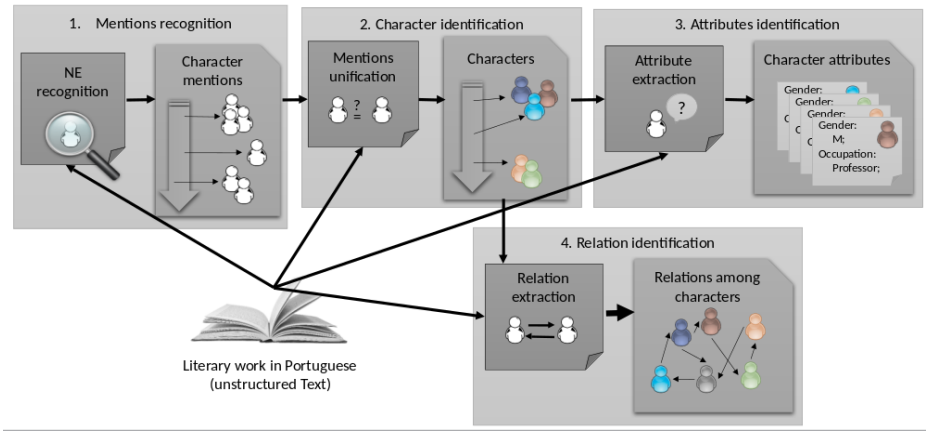

Fig. 2. Different subtasks in DIP (figure adapted from [4])

only plot characters. Historical people, or characters from other literature should not be flagged as characters of a given novel. However, it is possible that historical figures appear as fictional characters – as is the case of *D. Pedro II* in *Dom Casmurro*. Note also that novelists always pressupose some shared knowledge by their readers, but this knowledge can be epochal, and/or regional, as the following example illustrates: *A criança ainda vai ficar mais famosa que a Catarina Eufémia*. Although DIP participants do not have to decide whether the book is talking about a historical person or a famous literary character in another work, one might need to access some knowledge base not to mark *Catarina Eufémia* as a character.

As one character is rarely always called the same way throughout a book, we have the widely known coreference problem. This is what the unification step is supposed to solve, identifying all mentions that refer to the same character. Mentions are often depending on the context: depending on who is talking or referring to her, the name may be radically different. So, a serious challenge is to identify the set of (proper noun) denominations by which a novel character is mentioned in a work. By just comparing the two example novels, we see that diminutives may refer to the same person (as in *Guida* for *Margarida*), or to another one named after another character (as in *Capituzinha* and *Capitu*), and that in both books there are characters with the same first name (*João* and *Ezequiel*).

The two last steps presented in Figure 2, attributes and relation identifications, aim at recognizing the gender, profession/occupation/social status of the detected characters, and the relationship between them. DIP allows the creation of character networks, currently a hot task, as can be seen in the overview by [4]. These can in turn be used for genre prediction, (visual) text summarization, comparing fiction with (social) reality, comparing different literatures, and deciding who are the main characters.

Other applications do not rely on networks, but concern simply the description of large amounts of data: the role of women and men as characters, the professions or social statuses mentioned, the most common relationships found. And it is conceivable that the kind of relations and address forms differ depending on the time of the plot and on the time of the book creation, providing clues for periodization and genre.

The expected results of DIP, are namely, per literary work (see Table 1)

1. a list of characters, each character represented by a list of possible mentions
2. the gender of the character (M, F, or M and F)
3. the profession, or occupation, or social status of the character (can be more than one, or none)
4. the family relations among any characters

2 Previous work

There are a number of works on automatic character recognition, see e.g. [1, 12, 2]. [6] identify characters in Portuguese children's books to attribute direct speech, while [9] use rules to create character networks for some literary works, distinguishing between plot characters and other named people who are either historical, or characters from other works.

As to named entity recognition in Portuguese, there has been significant work in (among others) person name recognition in HAREM [10], and also some form of identification among different denominations in ReRELEM [3] a decade ago. As to distant reading in Portuguese, cf. [7].

3 The DIP setup

DIP organization will provide 200 books in digital form (half in text, half in pdf format) to the community, which has 48 hours to return the results. This effectively prevents a close reading of the 200 works, ensuring that the analysis is done automatically.

In order to have large numbers of works to process, distant reading of literary collections necessarily includes works from several time periods – after all, one of the goals of distant reading is to address trends and changes in time, see [11]. This means that, specifically for Portuguese, systems will have to process several different ortographies and styles, including different ways to describe professions and relations. For example, *boticário* or *cacaolista* are not exactly modern words to refer to a pharmacist or a cocoa farmer, and there are few *fogueteiros* or *jograis* nowadays. The historical novel subgenre, quite frequent in Portuguese, brings a set of additional problems [8], such as old names, jobs and address forms.

After the submission period is over, the golden collection (containing the right information for 40 out of the 200 books) is made publicly available, and the evaluation results are computed. A workshop presenting the results and the

different approaches of the participants will then be organized, followed by the publication of a journal volume. All data amassed about the literary works will also be released.

The systems will be evaluated separately on the five tasks, with the final score per book being the sum of the five measures. The ranking among the systems is done by macro-averaging over the golden collection.

Finally, it was necessary to devise a relativey complex form for evaluating family relationships, inspired by [13], given that e.g. `X irmã de Y` (X sister of Y) and `Y filho de Z` (Y son of Z) conveys precisely the same information as `Z mãe de X` (Z mother of X) and `Y irmão de X`. (Y brother of X).

References

1. Bamman, D., Popat, S., Shen, S.: An annotated dataset of literary entities. In: Proc of NAACL 2019. p. 2138–2144 (2019)
2. Dekker, N., Kuhn, T., van Erp, M.: Evaluating named entity recognition tools for extracting social networks from novels. PeerJ Computer Science **5** (2019)
3. Freitas, C., Santos, D., Mota, C., Oliveira, H.G., Carvalho, P.: Detection of relations between named entities: report of a shared task. In: Proceedings of the NAACL HLT Workshop on Semantic Evaluations, SEW-2009. pp. 129–137 (2009)
4. Labatut, V., Bost, X.: Extraction and analysis of fictional character networks: A survey. ACM Computing Surveys **52**(5) (2019)
5. Langfeldt, M.C., Gaiotto de Moraes, R., Pires, E.C.: A importância do Desafio em Identificação de Personagens (DIP) para os estudos literários lusófonos. Tech. rep., DIP (2021), `https://www.linguateca.pt/aval_conjunta/dip/Langfeldtetal2021.pdf`
6. Mamede, N., Chaleira, P.: Character identification in children stories. In: Vicedo, J.L., Martínez-Barco, P., Muńoz, R., Noeda, M.S. (eds.) Advances in Natural Language Processing - EsTAL, pp. 82–90. Springer (2004)
7. Santos, D., Alves, D., Amaro, R., Branco, I.A., Fialho, O., Freitas, C., Higuchi, S., Langfeldt, M., Lopes, J.M., dos Santos, A.L., Pires, E., Ramos, B., Sanches, D., Fuão, R.S., Pereira, P.S., Terra, P.: Leitura Distante em Português: resumo do primeiro encontro. Materialidades da Literatura **8**(1), 279–298 (2020)
8. Santos, D., Bick, E., Wlodek, M.: Avaliando entidades mencionadas na coleção ELTeC-por. Linguamática **12**(2), 29–49 (dezembro 2020)
9. Santos, D., Freitas, C.: Estudando personagens na literatura lusófona. In: STIL - Symposium in Information and Human Language Technology. pp. 48–52 (2019)
10. Santos, D., Seco, N., Cardoso, N., Vilela, R.: HAREM: An Advanced NER Evaluation Contest for Portuguese. In: Calzolari, N.e.a. (ed.) Proceedings of LREC 2006). pp. 1986–1991 (2006)
11. Underwood, T.: Distant Horizons: Digital Evidence and Literary Change. University of Chicago Press (2019)
12. Valaa, H., Jurgens, D., Piper, A., Ruths, D.: Mr. Bennet, his coachman, and the Archbishop walk into a bar but only one of them gets recognized: On The Difficulty of Detecting Characters in Literary Texts. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. p. 769–774 (2015)
13. Vilain, M., Burger, J., Aberdeen, J., Connolly, D., Hirschman, L.: A model-theoretic coreference scoring scheme. In: Proceedings of the 6th Message Understanding Conference (MUC-6). pp. 45–52. Morgan Kaufmann (6-8 November 1995)

Table 1. Results for *As pupilas do Senhor Reitor* and *Dom Casmurro* (incomplete).

Characters	sex	profession
Margarida, Guida, Guida dos Meadas	F	cabreira, professora
Clara, Clarinha, Clarita, Clarita dos Meadas	F	
Daniel, Sr. Daniel, Danielzinho, Daniel do Dornas, Danielzinho do Dornas	M	médico, estudante
Francisca, Chica, Chica da Esquina	F	
Joana, Sra. Joana	F	criada
José das Dornas, Sr. José, Sr. José das Dornas, José, Sr. Zé, Sr. José	F	lavrador
Pedro, Pedro das Dornas, Sr. Pedrinho	M	lavrador
Sr. Reitor, Sr. Padre António, Padre António, Senhor Reitor	M	padre
João Semana, Sr. João Semana, João da Semana	M	médico
João da Esquina, Sr. João da Esquina, Sr. João, João	M	boticário
Sra. Teresa, Sra. Teresa de Jesus	F	
Zefa da Graça, Josefa da Graça, Zefa	F	
Álvaro, Sr. Álvaro	M	
Margarida irmã de Clara		
Daniel irmão de Pedro		
José das Dornas pai de Daniel		
João da Esquina marido de Sra. Teresa		
João da Esquina pai de Francisca		
Doutor João da Costa, João da Costa, Pai João	M	
Cosme, Mano Cosme, Primo Cosme, Tio Cosme	M	advogado
Ezequiel A. de Santiago	M	
D. Glória, Dona Glória, D. Maria da Glória Fernandes Santiago, Prima Glória, mana Glória	F	
Pedro de Albuquerque Santiago	M	fazendeiro, deputado
Sancha, Sanchinha, D. Sancha, Sinhazinha Sancha, sinhazinha Gurgel	F	
Justina, Prima Justina, D. Justina	F	
Padre Cabral, Cabral	M	padre
Pádua, João, Sr. Pádua, Joãozinho, Tartaruga	M	funcionário público
Dona Fortunata	F	
Bento, Padre Bentinho, Sr. Bentinho, Doutor Santiago, Dom Casmurro	M	
Escobar, Ezequiel de Sousa Escobar	M	investidor em café
José Dias, Sr. José Dias	M	médico, agregado
Capitu, Capitolina	F	
Capituzinha	F	
Miquelina	F	escrava
Maria Gorda	F	escrava
D. Pedro II	M	imperador do Brasil
Bento marido de Capitu		
Capitu mãe de Ezequiel A. Santiago		
D. Glória viúva de Pedro Santiago		
D. Glória mãe de Bento		
D. Fortunata mãe de Capitu		
D. Fortunata esposa de Pádua		