jcls. Journal of Computational Literary Studies

OPEN ACCESS

Article

Gender Depiction in Portuguese
Distant Reading Brazilian and Portuguese Literature

Cláudia Freitas[1]
Diana Santos[2]

1. Department of Letters, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, Brazil.
2. Department of Literature, Area Studies and European Languages, University of Oslo, Oslo, Norway.

Citation
Cláudia Freitas and Diana Santos (2023). "Gender Depiction in Portuguese. Distant Reading Brazilian and Portuguese Literature". In: *Journal of Computational Literary Studies* 2 (1). 10.48694/jcls.3567

Date published 2024-02-14
Date accepted 2023-12-11
Date received 2023-01-19

Keywords
distant reading, annotation, Brazilian literature, Portuguese literature

Reviewers
Ulrike Henny-Krahmer, Mareike Schuhmacher

Note
This paper has been submitted to the conference track of JCLS. It has been peer reviewed and accepted for presentation and discussion at the 2nd Annual Conference of Computational Literary Studies at Würzburg University in June 2023.

Abstract. In this paper, we look at how masculine and feminine characters are described in literature in Portuguese using a publicly available literary corpus: *Literateca*. We investigate the words used to characterise human beings, after classifying them into four broad categories, namely those related to the social, appearance, character and emotional axes. We study the influence of genre, literary school, author gender, and time, among others.

1. Introduction

The way people are described is a rich source of information about societies and cultures, revealing the values and beliefs of those who describe them. In addition to proper names, there are many other ways of human designation, such as the use of general human nouns like *man, woman, person, gentleman, lady*, and designation by traits or functions of the people mentioned (using places of origin, professions, family ties, etc., such as *Brazilian, doctor, mother, foreigner*).

In this paper, we look into how human beings are characterised in literature in Portuguese – also called Lusophone literature – using a distant reading approach. In particular, we want to investigate the influence of features such as authorship, geographical origin, historical period and gender (both character gender, and authorial gender).

Inspired by Moretti and Sobchuk (2019)'s warning, we try to go beyond simple visualisations by date or author, and add other ways to look at the data. Following their "dissecting table" analogy, our aim is to find out which pieces are able to provide pertinent analysis, triggering meaningful readings. So, we search for "creative cuttings", – such as the "volume" of speech verbs in Katsma (2018) – to give us new insights. Specifically, we add the class 'human depiction' to our data; still, we aim for consensual and understandable categories, like "century" in history.

1.1 Gender in Literature

The theme of gender roles in fiction texts has received increasing attention in the Digital Humanities community, as the following works testify.

Looking at English literature (104,000 works, from 1703 to 2009), Underwood et al. (2018) found that the gender difference between characters became less pronounced from the middle of the nineteenth century to the present day: Actions and attributes of characters became less defined by gender categories. In other words, gender roles tend to become more flexible. At the same time, they also found a decrease in the number of feminine characters, with the volume of fiction written by women from 1850 to 1950 dropping by half.

Exploring the *Black Drama* collection, which contains plays written between 1950 and 2006, Argamon et al. (2009) report poor results when trying to automatically distinguish the gender of the authors and/or characters. However, they found differences in the way masculine and feminine authors and characters use language. Feminine playwrights allocate more than half (52.1%) of speeches to feminine characters, while 34.7% of the speeches in plays by masculine authors belong to feminine characters.

Working with present-day Dutch literary fiction (170 novels published in one sample year), Smeets (2021) found the same imbalance between masculine and feminine characters. However, the author questions what he describes as a "perhaps naive mimetic assumption" according to which the relative absence of feminine characters is a result of their unequal status in society. From the results of his investigation, feminine characters, although fewer in number, occupy a relatively central position in their fictional social networks — they display more relations, both more relations in general and more relations with important characters.

Hoyle et al. (2019), using 3,5 Mio. digitised books in English, analyses the lexical choices (adjectives and verbs) associated with feminine gendered nouns and found that positive adjectives used to describe women were more often related to their bodies than adjectives used to describe men. Following the same trend, Schulz and Bahník (2019) explores the depiction of male and female characters using the Google Books Ngram corpus, focusing on twentieth-century English-language fiction. The study analyses adjective-noun bigrams associated with the words *man*, *woman*, *boy*, and *girl*, and reports that adjectives associated with *men* are more positive ("honest", "wise", "honorable", and "able") than those associated with *women* ("vulgar", "foolish"). As for preferences, "charming", "fashionable", and "warm" were relatively feminine words, while "lazy" and "mean" were relatively masculine words. On the one hand, men were described in decreasingly masculine terms throughout the beginning and end of the twentieth century; on the other hand, the masculinity of adjectives used to describe women started to slightly increase from 1968 to 2000.

Weingart and Jorgensen (2013) performed a computational analysis of gendered bodies in ca. 200 European fairy tales (German, French and Italian folklore texts translated into English). They show that feminine characters are more likely than masculine characters to be described with appearance-evaluative words, suggesting that men are associated with the mind and women with the body.

Cermáková and Mahlberg (2022) explore linguistic descriptions of gendered body language and compare nineteenth century British children's literature (*ChiLit Corpus*) with contemporary fiction for children (the *OCC2000+ Corpus*, a subcorpus of the *Oxford Children's Corpus*). Using a corpus linguistic approach, the authors study sequences of five words which contain at least one body part noun and a marker of gender. They found fewer clusters for feminine characters in the nineteenth century. The contemporary data suggest a trend for feminine and masculine clusters to become more similar, and an increasing range of options for the description of feminine characters and their interactional spaces. Using the same *ChiLit Corpus*, Cermáková and Mahlberg (2021) focused on nouns — excluding proper names — frequently used to label people, and found that *Mothers* are the most frequent occurring feminine character in the corpus.

It is also worth noting the existence of studies such as Cao and Daumé (2021) and Lucy and Bamman (2021). The first one explores the consequences of gender bias for machine learning. The paper investigates how different aspects of linguistic notions of gender impact an annotator's judgements of anaphora, and points out that a significant possible source of bias comes from the annotations themselves — from underspecified annotation guidelines and the human annotators. The authors emphasise that both, humans and systems, should not over-rely on cues such as names, semantically gendered nouns, and terms of address, relying on "relatively safe" cues like syntax instead. At the other pole of the machine learning approach, the study conducted by Lucy and Bamman (2021) raises questions on how to avoid unintended social biases when using large language models for storytelling. Focusing on how GPT-3 may perceive a character's gender based on textual features such as personal pronouns (*he*/*she*/*her*, etc.), the work finds that stories generated by GPT-3 place masculine and feminine characters in different topics and exhibit many gender stereotypes: For example, feminine characters are more associated with family and appearance than masculine characters.

In this paper, we also try to contribute to the investigation of gender roles using works written in Portuguese. As a crossover between Corpus Linguistics and Digital Humanities, we use morpho-syntactic and semantic information automatically provided by the PALAVRAS parser (Bick 2014), and add extra semantic annotations, which are described below.

With Larson (2017), we recognise that using gender as a variable in Natural Language Processing is an ethical issue and that we need to explicitly explain what "gender" means in this work. As Larson points out, there are many views of how gender functions as a social construct. In this study, we treat gender as binary, since in the vast majority of works in our corpus, gender was mainly constructed in terms of the binary distinction femininity/masculinity. We acknowledge, however, that the category "gender" can be more complex than this binary distinction, and that these kinds of studies, which describe the cultural apparatus around gender for an extended period of time, do not in any way purport to assert what gender is, but only how it has been/is perceived. So they should not be used for reinforcing gender stereotypes, as warned against by Mandell (2019).

1.2 Previous Work for Portuguese

For distant reading of Portuguese, we are aware of some works dealing with characters in literature (Santos and Freitas 2019), as well as of the DIP challenge for automatic character identification in Portuguese (Santos et al. 2022b), to which we will come back later.

Our point of departure is the work by Freitas et al. (2022)[1] – and later extended in Silva's master thesis (Silva 2021) – who have suggested a fourfold classification for human

[1] Although published in 2022, the work was conducted in 2018.

characterisation. Human attributes were organised in social, appearance, character, and emotional characteristics.

Using *OBras*, a corpus of Brazilian literature in the public domain (Santos et al. 2018), they studied 223 works by 25 Brazilian authors, two of them women (authoring 3 novels altogether), and observed the following trends:

- Men were more frequently described than women (60%-40%), which may be related to the fact that there were roughly more masculine characters than feminine ones in the same proportion.
- The most frequent masculine characterising words were *bom* (good), *sério* ('honest'), *rico* ('rich'), and *alto* ('tall'), while *bonita* ('beautiful') was by far the top characteristic for women.
- Almost 50% of women depicting words were about beauty (namely *bonita* and *bela*).
- Character and social predication were most frequent for men; for women, social characterisation is reduced to *married* and *rich*.
- Emotional characterisations like *feliz* ('happy') were (almost) exclusively used for women.

We wanted to check whether these observations held true for a wider collection, including Portuguese literature as well.

1.3 A Brief Comparison with DIP

It is useful to compare and contrast our study with the recent DIP challenge for Portuguese (*Desafio de Identificação de Personagens*), an evaluation contest for identifying literary characters and some information about them in Brazilian and Portuguese works (Santos et al. 2022a, 2023). By describing them and pointing out the differences, we shed some light on different ways of looking at (roughly) the same data.

For DIP, the unit is the literary character, and so the challenge looked at their gender, their profession, occupation and/or social status, and their family relationships with other characters. In addition, "literary character" in DIP does not include all people.[2] In the present study, we try to look at all mentions of characterisation of people in the works, so all numbers reported in this paper are not per character, but per mention of people.

We will discuss and compare the findings about character gender in subsection 4.7.

1.4 The Importance of Studying Literature in Portuguese

Portuguese has a rich literary tradition, but unfortunately the digitisation efforts are lagging behind other languages. This has, for example, been discussed in Schöch et al. (2021).

[2] By this, we mean that when people are mentioned to specify a time frame or authorship, as in *During D. João VI's reign*, or as in *Goethe's Faust*, neither "D. João VI" nor "Goethe" were considered characters. But this turned out to be a controversial decision and hard to decide in historical novels. In any case, it does represent an unusual way to look at literary characters that needs to be documented.

Also, major actors in the big data landscape, no matter the high number of Portuguese speakers in the world, have not endowed Portuguese with the "current" tools that are available for other languages, even with much fewer speakers/readers/writers, like Hebrew or Italian: There is, for example, no Google Book N-grams[3] service for Portuguese.

Likewise, recent reviews of the computational literature landscape, because they do not find enough internationally published DH papers on Portuguese, have decided not to review or include papers in Portuguese, therefore contributing actively to the lack of information on Lusophone materials and studies. For example, Schöch et al. (2022, 4) state:

> "several languages, however, were represented only with relatively low numbers of articles or papers, and in order not to misrepresent the research communities these publications stem from, we decided not to take the materials in several languages into account [...]."

This is one of the reasons why we are writing this paper for an international audience. Maybe the results are not so different than the ones our English-speaking or English-studying colleagues obtained, but they are novel because they are obtained from completely different data.

2. The Material

We provide here an overview of the data used, also with the purpose of making it known, and hopefully, useful, for other researchers. And not least because it shows the methodological problems it invites.

Attempting to complement close readings of canonical authors with a wider material, following Moretti (2000, 2013) and Underwood (2019), we use as many books as possible whose full text is currently publicly available in Portuguese to investigate properties of literary text which can be identified in an automatic way.

In order for these data to be shareable and replicable for studies, we restrain our data (mostly[4]) to books in the public domain. We are aware that many more texts exist in electronic form, but by using them we would either incur copyright law infringement, or at least we would risk creating materials only for our own study, which cannot be shared with others.

Also, it is important to stress that we are referring to textual versions of the works, not simply images. Optical character recognition for Portuguese, especially for old books, is not good enough yet, so all books have been revised by humans, if not born digital.

2.1 Corpus

We used *Literateca* version 11.1, created on 26 May 2023, comprising ca. 32 Mio. tokens of (original) prose (excluding drama) from 1700 on. A quantitative overview of the

[3] See: `https://books.google.com/ngrams/`.

[4] Exceptions are excerpts of books existing in parallel corpora or texts whose authors gave us permission to use them.

Literature	no. of tokens	no. works	no. authors
Total	32,718,621	669	200
Portuguese	20,639,007	306	127
Brazilian	12,079,614	355	73

Table 1: Size of the material: prose from 1700 to the present.

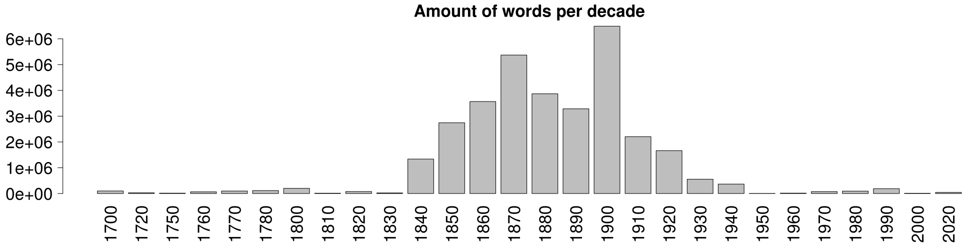

Figure 1: Distribution of words per decade.

material is in Table 1. Figure 1 shows the distribution of the material in time, by size in words.

Literateca is the result of the merging of several literary corpora written in Portuguese, and thus has some particularities:

- It includes literary works by canonical authors, but also other works by those canonical writers that are not usually or necessarily deemed literary, such as newspaper chronicles, letters, memoirs, and even scholarly works such as history books or ethnographic studies, and travelogues. For earlier centuries, even sermons are included. However, these genres are only included for canonical writers.[5]
- It includes drama, poetry, and prose.
- Some of the works have updated orthography, others keep the original orthography. Given that there have been several norms of Portuguese spelling across the centuries, this means that there can be a variety of forms for the same word.
- While some authors have all their works included, others have only a few, or just one. Especially for non-canonical writers, there is no claim to completeness.
- Given that the works have been digitised by different bodies and with different tools and for different purposes, there is no claim to homogeneity: Works can come from the first or the last paper version, they may keep their prefaces or not, they have different ways of describing chapters, etc.
- All works are marked with author, author gender, date of publication, variety of Portuguese, genre, and whether they are original or translated. Some texts are also classified by the literary school they belong to.

We tried to use as much of this material as we could, but we removed poetry and drama. Poetry is probably a natural choice to remove because of syntactic idiosyncrasies – and therefore a worse parser performance –, and because we believe that poetry has not so

[5] By this, we mean that established authors who belong to the Portuguese and Brazilian canons have been fully digitised, i.e., everything they published is available. This is in strong contrast with the works of non-canonical authors, who may have had some of their (mainly) novels digitised in the context of other projects.

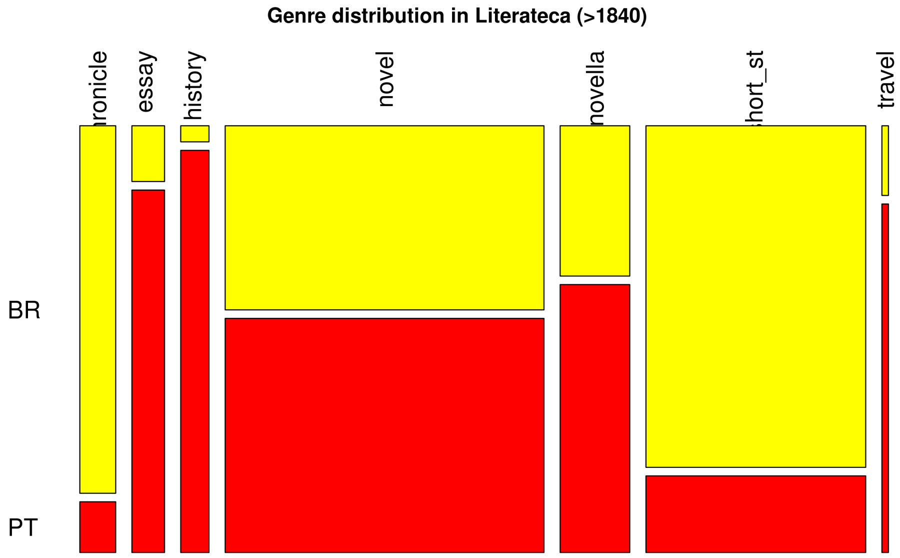

Figure 2: Genre in the full corpus. The unit is the work. In red are the works written by Portuguese authors.

	Fiction	Non fiction	Total
Brazil	10,547,327	1,532,287	12,079,614
Portugal	15,280,938	5,358,069	20,639,007
Total	25,828,265	6,890,356	32,718,621

Table 2: Size in words of the different materials after 1840.

many mentions of fictional characters. We removed drama, also in prose, because it was heavily unbalanced, given that most of the plays were from Portugal.

As to prose, we started to use everything published since 1700. It is, anyway, important to recognise that we do not have a balanced corpus, and the lion's share is fiction. We then selected different subsets for different research questions.

- Just fiction and just non-fiction, to see whether the character depiction was different across the fiction divide.
- Just works published after 1840 to compare Brazilian and Portuguese authors.
- Just fiction published after 1840 to compare Brazilian and Portuguese literature.

See Figure 2 and Figure 3 for a bird's eye view of the genre distributions in total and in fiction.

Only in Figure 3, we include the variable author gender, since it is only in fiction that we have texts written by women.

In Table 2, we give the number of words in the material published after 1840.

2.2 Gender Attribution

We explore the influence of gender in both character description and authorship. Masculine and feminine gender labels were manually ascribed to writers, as our corpus contains

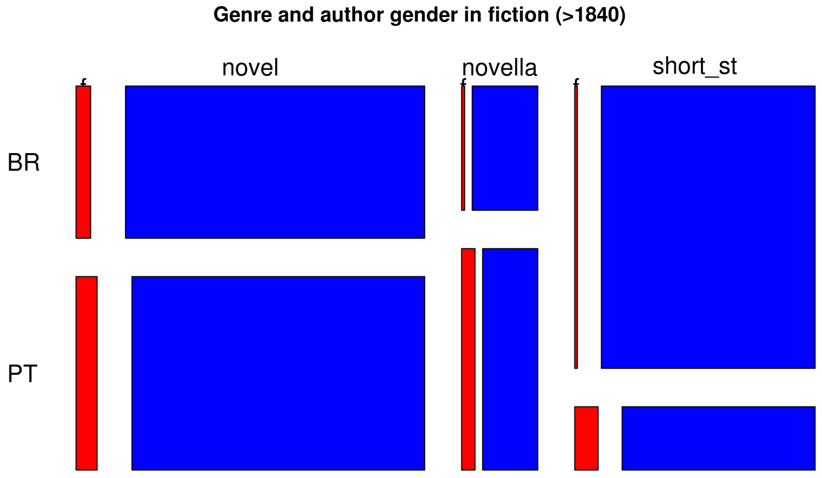

Figure 3: Genre in the fiction corpus. The unit is the work. In red are the works written by feminine authors.

works written by canonical authors, which are widely discussed in literary studies. For the non-canonical authors, gender was attributed either based on adjective/inflected forms used in prefaces or based on their proper names. As for the characters, the gender labels were automatically assigned by the PALAVRAS parser, and then manually revised by linguists (Rocha et al. 2019; Silva 2021). The linguistic clues that were followed on attributing and revising gender were syntactic agreement and morphological features.

Portuguese is a Romance language that forces the speakers to specify the gender of nouns (both common and proper nouns) and adjectives. The main formal clue to distinguish masculine and feminine forms is the word's ending: Masculine forms tend to end in *-o*, feminine ones tend to end in *-a*, and those ending in *-e* can be both feminine and masculine – *ponte* ('bridge') is feminine, and *pente* ('comb') is masculine. However, there is no perfect equivalence between the ending in *-o* or *-a* and the masculine or feminine gender, respectively – *planeta* ('planet') is masculine, and *tribo* ('tribe') is feminine. Therefore, observing syntactic agreement between the head noun and its modifiers is the most reliable way to assign morphological gender.

When calculating the gender of depicting words, we take into account the gender of the nominal head (noun, proper noun or pronoun) being characterised, not the gender of the words (modifiers) associated with it. This choice is due to the fact that, although adjectives can be inflected for gender in most cases, the search patterns we used also retrieve nouns, which do not admit inflection. Thus, nouns like *anjo* ('angel') will always be masculine, even if the mentioned angels are feminine. When considering the gender of nominal heads, *anjo*, although a masculine common noun, is classified as a feminine classifier if it modifies a feminine character.

3. The Process

We wanted to identify all cases where human beings were mentioned to find out how they were described or depicted. We extended the search patterns used by Silva (2021)[6] in two ways: (i) We enriched the lexicon of general human nouns, including names of professions as targets, and (ii) having extended the number of works analysed to include works written by Portuguese authors, we broadened the lexicon of characterising words[7], based on the prose of the eighteenth, nineteenth and twentieth centuries in *Literateca*. During the process of data analysis, we were forced to discuss the previous classification, which led to a refinement of the classification guidelines and a reclassification of a few words.

We start from the idea that specific linguistic patterns indicate certain (semantic) relationships. So, we have used a set of patterns – relying on the automatic morpho-syntactic annotation – to search the material for instances of describing human beings. Below are some examples of what the patterns yielded (the patterns are publicly available).

(1) – Ouviste? – perguntou **ela inquieta**. [– Did you hear? **she** asked **restlessly**.]

(2) ...acudiu logo o **padre**, muito **arisco**. [...came the **priest**, very **skittish**.]

(3) Uma **mulher honesta** não tem segredos para seu marido! [A **honest woman** has no secrets from her husband!]

(4) **D. Joana Tecla** era **idiota**. [–**Mrs. Joan Tecla** was an **idiot**.]

(5) Em todo o caso era uma bela **mulher**, **alta** e forte sem ser gorda... [In any case, she was a beautiful **woman**, **tall** and strong without being fat...]

(6) ...calado como a tarde triste, um **homem**, ainda **moço**, vestido como os essênios taciturnos, caminhava... [...silent as the sad afternoon, a **man**, still **young**, dressed like...]

Then we proceeded to classify each word of the aforementioned list – which are the words associated with human beings in the examples –, in four (non-mutually exclusive) classes, according to the type of characterisation: social, emotional, physical (appearance), and character. In order to group these idiosyncratic data and provide a better view from afar, we analysed the most frequent words and came up with the four classes. We also used the class 'other' if none of the four could hold, and one or more of the four otherwise. The allocation of the categories themselves follows their scope and the main choices involved are:

social In addition to professions, occupations, and social status, we also included absence of profession like *mendigo* ('beggar'), nationality, civil status, family relations, political opinions like *monárquico* ('monarchist'), and cases which are a

[6] Which, in turn, are an improvement of the patterns used in Freitas et al. (2022).

[7] The list comprises not only adjectives and nouns, but also verbs (for past participles), given that it is a feature of PALAVRAS that most participles are analysed as verbs even though in an adjectival context.

consequence of social intercourse, like *ignorante* ('ignorant') or *educado* ('civil' or 'knowledgeable').

appearance Physical appearance, including clothing or lack of it, as well as those features associated with time, such as *jovem* ('young') or *velho* ('old').

emotional Feelings, emotions, and emotional tendencies.

character Personality traits, also including cognitive properties, such as intelligence or lack of it. It also includes evaluations according to social conduct, such as *honesto* ('honest'), *malcriado* ('rude') or *pretensioso* ('snobbish').

It is important to mention that each category works as a label, which in turn encodes four perspectives on people: 'Appearance' refers to what is visible; 'social' refers to the various roles someone can play in society; 'character' refers to internal/cognitive characteristics; and 'emotion' refers to emotional traits. We could also, and more broadly, consider two large classes: internal characteristics ('character' + 'emotion') and external characteristics ('appearance' + 'social'). We will use this in Figure 17 below. We note that the words classified can often refer to non-human entities, as in the next example (7). But if they could modify a human person, they were classified accordingly. However, the results presented in the next sections refer only to those cases where the characterisation was assigned to human beings, as in example (8), since only they are retrieved by the patterns applied.

(7) – Que **triste** pensamento! [What a **sad** thought!]

(8) – Mas a **triste** senhora continuava a choramingar. [But the **sad** woman kept weeping.]

We classified the retrieved words out of context, except in those rare cases where we had to check whether the adjective had been used to characterise at all in the corpus.[8] For example, initially, we wanted to discard the words *granítico* ('made of granite') and *triunfal* ('of triumph'), but we checked the corpus and there were instances where both were applied to human characters, so they were retained in our list.

(9) – Sim, o velho Afonso é **granítico**... [– Yes, old Afonso is **made of granite**...]

(10) Nunca as mulheres **triunfais** me fizeram bater o coração... [**Triumphal** women never made my heart beat...]

The classification was done manually by the authors of this paper, and divergences were heartily discussed. We dismissed mistakes either because (i) they were not characterisation words, (ii) they resulted from wrong parsing, or (iii) we decided they were not relevant to our goals. As for exclusion:

[8] Actually, there was one case where we consistently considered the context: In Portuguese, the word *grande* can mean either *big* or *great*. Since each meaning corresponds, in general, to a different syntactic position – *grande homem* ('great man'); *homem grande* ('big man'), we used this information to correctly classify each of the occurrences: 'character' or 'appearance', respectively.

type	size
Social	1,391
Appearance	672
Emotion	514
Character	1,578
Other	326
Total	4,481

Table 3: Depicting words by category. Recall that words can belong to more than one category.

- We did not take into account "complex adjectives" in the sense of having more than one word, like *bem intencionado* ('having good intentions'), *mal intencionado* ('having bad intentions'), *bem educado*[9] ('polite'), etc.

- We did not classify relational adjectives, such as *partidário* (*de...*) ('partisan'), *apologista* (*de...*) ('in favour of'), *comparável* (*a ...*) ('comparable to'), *emparelhado com* ('pairing with'), *similhante a* ('similar to'), since a precise characterisation would require a close reading of each sentence.

- We dismissed misspellings, except for lack of diacritics.[10] Our rationale is that, in future improved versions of the corpus, the corrected words would be correctly annotated.

Following the annotation approach adopted in the AC/DC project (Santos 2014), which underlies *Literateca*, we used multiple classification when two or more categories/senses could be assigned to a characterising word (vague or ambiguous words). References to madness, for instance, were considered both 'social' and 'character'. The same is true for habits like *madrugador* ('early riser') and *bêbado* ('drunkard' or 'drunk'), which can be either due to biology or social upbringing. The word *acanhado* ('shy') can be interpreted as a not-expansive person (thus 'character') or as someone fearful ('emotion'), and the same applies to *impaciente* ('impacient'), which can be interpreted as anxious ('emotion') or restless ('character').

Finally, cases such as *maravilhoso* ('wonderful'), *incomparável* ('incomparable'), *ideal* ('ideal') or *horrível* ('horrible'), where it is not clear to which axis they apply out of context, were classified as referring simultaneously to 'character', 'social' and 'appearance'.

To verify the degree of reliability of the classifications and the adequacy of the classes, Silva (2021) carried out a study on the inter-annotator agreement of 15 people in the classification of occurrences considered especially difficult. The degree of agreement was 80%. We have not carried out any further studies on this matter.

After this classification, we ended up with a list of 4,481 words which might be employed in depicting human beings (see Table 3).[11] Due to the properties of the parser, we list the lemmas which can be verb infinitives for past participle forms, because we use the lemmas in our patterns.

[9] But note that *educado* and *bem-educado*, as words of size one, were included.
[10] That is, we considered missing accents to be something that could be present in the original paper edition but not OCR mistakes.
[11] Available from `https://www.linguateca.pt/Gramateca/ListaPredicadoresClassificados.txt`.

type	size
Appearance, character	88
Appearance, emotion	12
Appearance, social	8
Appearance, character, social	10
Character, emotion	107
Character, emotion, social	1
Character, social	80
Emotion, social	9
Total with more than one class	315

Table 4: Words belonging to several categories.

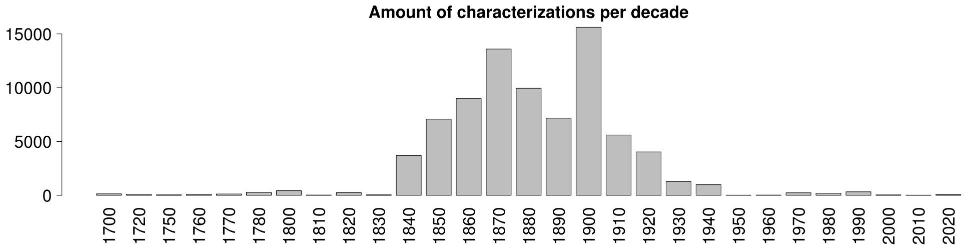

Figure 4: Number of mentions of characterised people in the corpus per decade.

In order to provide a richer description of this list, we show in Table 4 how often depicting words are vague or ambiguous.

We then annotated the corpus with this new classification[12] and computed how often and when the words were used to describe human beings.

We start by providing a picture of the distribution of mentions of human characters over time in Figure 4, as well as how many depicting events we were able to identify in Figure 5.

A comment is in order: The decade of 1830 is a clear outlier because it contains only one short text of 19,334 words, a political pamphlet by Alexandre Herculano, in the whole decade. The same happens with 1950, which is represented in the material by only 4,777 words of Jorge Amado's *Gabriela, Cravo e Canela*.

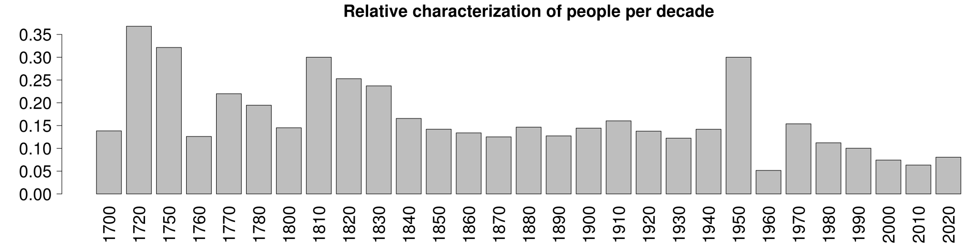

Figure 5: Relative characterisation per person, per decade.

[12] The classification is encoded in the following tags `pred:carater`, `pred:aparencia`, `pred:social` and `pred:emo`. To find them in *Literateca*, search for `[sema=".*pred:social.*"]`, etc.

4. Analysis

The first thing we report is the proportion of these subclasses in our material. Table 5 shows the raw numbers, and also those referring to masculine and feminine characters.[13] Figure 6 displays the overall distribution of characterisation words.

	Total	Mentions of masculine characters	Mentions of feminine characters
People	578,815	352,851	173,370
Characterised people	80,415	52,252	24,664
Social	11,793	7,813	3,534
Appearance	15,394	9,099	5,862
Emotion	9,670	5,562	3,895
Character	23,880	16,542	6,394

Table 5: Different depiction classes, in general and per gender of the characterised person, using the subject's gender.

The first observation is that there are many more mentions of masculine than of feminine characters in the material (ca. twice as many). Feminine characters are, however, almost as often characterised as the masculine ones: 14.2% against 14.8%.

The second remark is that by far the most frequent subclass deals with character (most frequent words: *bom* ('good'), *grande* ('great'), *honrado* ('honourable, honest'), *simples* ('simple'), *digno* ('with dignity'), *excelente* ('excellent')), followed by appearance (most frequent words: *velho* ('old'), *novo* ('young')[14], *antigo* ('old-fashioned'), *jovem* ('young'), *belo* ('beautiful'), *formoso* ('beautiful'), *bonito* ('pretty')).

Social characterisation comes third (most frequent words: *rico* ('rich'), *ilustre* ('illustrious'), *nobre* ('noble'), *casado* ('married'), *célebre* ('famous'), *pobre* ('poor'), *livre* ('free'), *famoso* ('famous'), while emotional characterisation is the least frequent (*pobre* ('poor'), *infeliz* ('unhappy'), *valente* ('brave'), *feliz* ('happy'), *triste* ('sad'), *desgraçado* ('miserable), *alegre* ('joyful'), *humilde* ('humble')).

Thirdly, feminine characters have a higher chance of being characterised by their appearance compared to masculine ones (23.8% vs. 17.4%), which confirms the findings of previous studies, and which we return to in subsection 4.2.

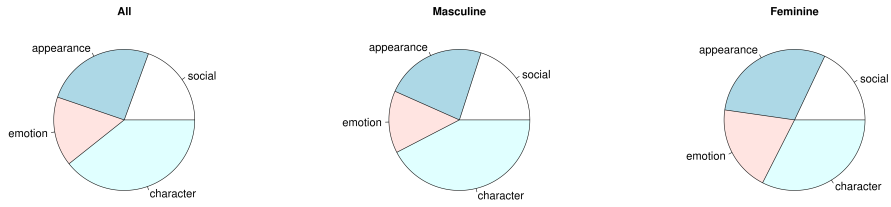

Figure 6: Distribution of characterisation words among the four classes for all, masculine and feminine, depictions.

[13] It should be noted that the numbers do not add up because in some cases the parser is not able to assign a morphological gender and marks them as M/F. Also, remember that by "character" here we mean mentions to people, not distinct characters.
[14] It may seem surprising at first to include age as appearance, but it is something that we assess visually.

	Total	Masculine	Feminine
Words	25,828,265		
Mentions of people	490,892	291,403	159,216
Characterised mentions of people	47,450	30,036	16,620
Social	8,968	5,720	2,979
Appearance	12,951	7,401	5,226
Emotion	8,767	4,922	3,665
Character	19,002	12,587	5,773

Table 6: Different depiction classes, in general and per gender of the characterised person, using the subject's gender only in novels, novellas, and short stories.

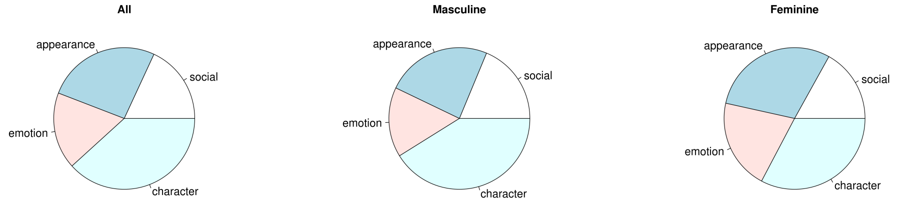

Figure 7: Relative characterisation per gender in novels, novellas, and short stories.

4.1 Does Textual Genre Matter?

Does it make more sense to look only at literary texts, removing travelogues, essays, history and political writings?

On the one hand, we kept all the material because we wanted to look at the way people described people in Portuguese, but then it is also conceivable that the kinds of information about people are rather different when you write the history of the Inquisition, an essay about your fellow writers, or a report of you crossing Africa, compared with a narrative in which you introduce fictional characters.

So, we reproduced our queries, removing all texts not classified as novels, novellas or short stories (see the new numbers in Table 6).

It is interesting to see that removing the non-fictional prose genres does not change the relative order of the subcategories but increases the percentage of feminine characters, from 30.0% to 32.4%, and characterised feminine characters, from 33.2% to 35.0%.

As to the characterisation of masculine and feminine characters, we have similar trends to those presented for the full material, as shown in Figure 7: Masculine targets are characterised, by far, by their character, while feminine targets are (almost) equally characterised by their appearance and their character.

For the non-fiction part, let us see whether the picture is different. In Table 7, we describe the masculine and feminine characterisations in the (considerably smaller) non-fiction part.

The percentage of feminine characters and feminine characterisations shrunk considerably to 16% and 18%, confirming that women are even less important in the public sphere.

	Total	Masculine	Feminine
Words	6,890,356		
Mentions of people	87,923	61,448	14,154
Characterised mentions of people	10,537	8,033	1,899
Social	2,825	2,093	555
Appearance	2,443	1,698	636
Emotion	966	688	245
Character	4,878	3,955	621

Table 7: Different depiction classes, in general and per gender of the characterised person, using the subject's gender only in non-fiction.

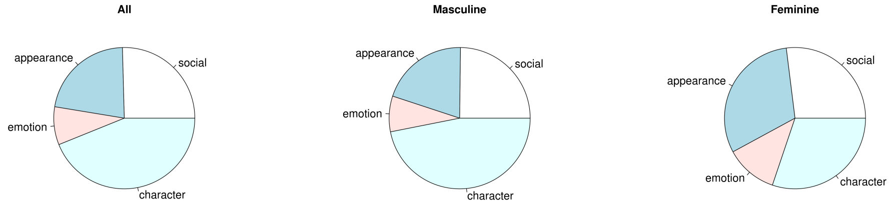

Figure 8: Relative characterisation per gender in non-fiction.

We see that social characteristics are – globally – more frequent than appearance. 'Character' remains the most important form of describing people, and 'emotion' the least.

In Figure 8, we present the distribution of the four kinds of features and see that the few mentions of women that are present have a large proportion of appearance descriptions, even more in non-fiction than in fiction.

4.2 Differences when Describing Masculine and Feminine Characters

The previous figures show that 'appearance' is more frequently used when describing feminine characters. Based on the entire data set, this can also be seen in the bar plot in Figure 9.

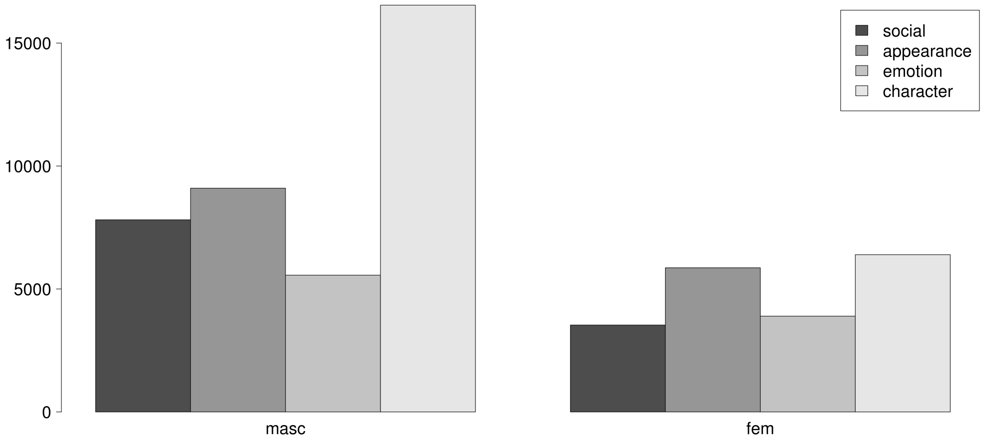

Figure 9: Relative characterisation per gender in the whole material as a bar plot.

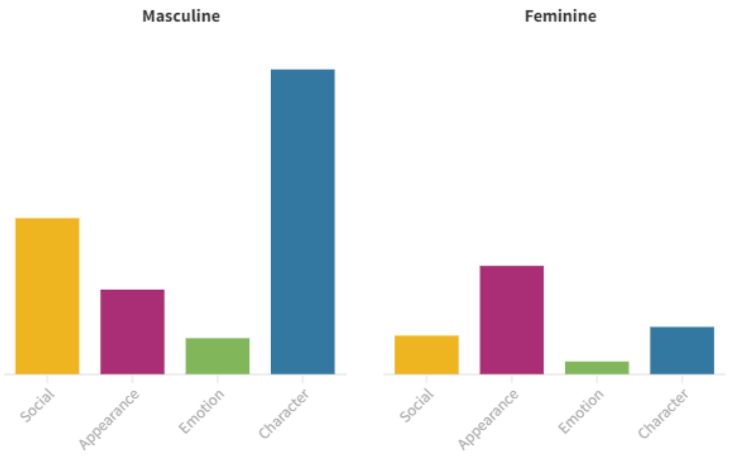

Figure 10: Preferred characterisation per gender.

However, this is just the tip of the iceberg. The analysis of depictive words preferentially used with masculine and feminine characters can be more revealing than the general analysis we presented in Figure 9, which takes into account the whole range of depictive words. In order to be evaluated as ‘preferred’, a word must (i) be used for masculine targets at least 80% of the occurrences, or for feminine targets more than 60% of the occurrences; and (ii) have a total frequency of 4 or more.

In cases where different lexical items correspond to gendered male/female pairs (*mãe/pai* (‘mother/father’); *rainha/rei* (‘queen/king’); *namorada/namorado* (‘girlfriend/boyfriend’) etc.), we manually grouped the elements of the pair as if they shared the same lemma, so that they could be included in the preference count.

The new data are presented in Figure 10, which shows a slightly different picture, in which (i) words of the emotional axis are almost not seen at all and almost disappear with the feminine characters, (ii) the balance between ‘appearance’ and ‘character’ in feminine depiction gives way to a characterisation based mainly on ‘appearance’, which accounts for half of all preferred feminine characterisations, and (iii) ‘appearance´, the second most frequent characterisation (of both masculine and feminine characters), drops to third place when associated with masculine characters, and rises to the first place when associated with feminine characters.

The ‘appearance’ axis has a raw frequency almost similar for both genders, but Figure 11 and Figure 12, complementary to Figure 10, provide a few details that enrich the analyses.[15]

As noted in previous studies, typically feminine social characterisations relate to the family environment (*mãe* (‘mother’), *prima* (‘cousin’)). However, mentions of the marital status are the highlight: (*casada* (‘married’) and *viúva* (‘widow’) are the most frequent words, but *adúltera* (‘adulteress’) is frequent as well. Conversely, marital status is absent as typical masculine social characterisation. They are rather related to (positive) social recognition such as *ilustre* (‘illustrious’), *célebre* (‘famous’), *notável* (‘remarkable’), *famoso*

[15] In Figure 11 and Figure 12, words such as *beautiful_1* and *pretty_2* relate to different Portuguese words that could be translated into the same English word, such as *bonita* e *formosa*, which could be both translated as ‘pretty’.

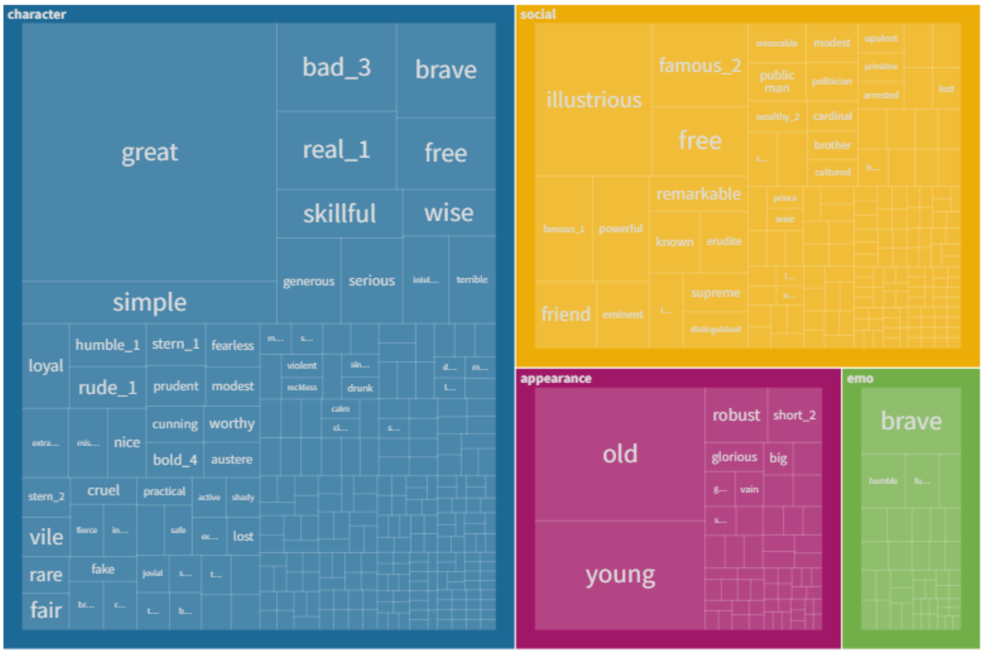

Figure 11: Preferred characterisation of masculine characters.

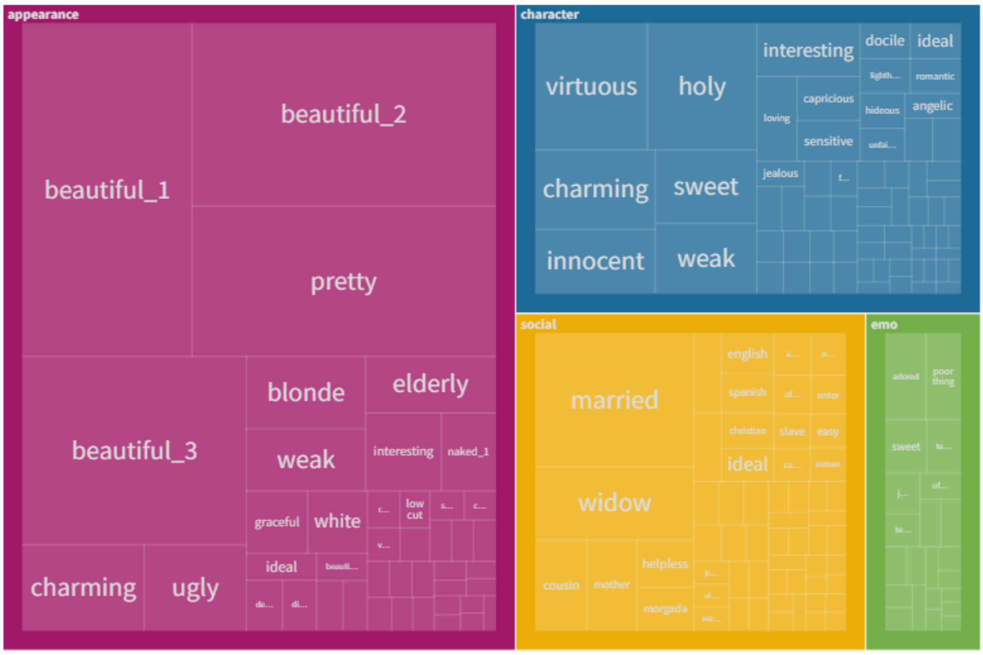

Figure 12: Preferred characterisation of feminine characters.

(another word for ‘famous’), and *poderoso* (‘powerful’).

On the feminine emotional axis, words associated with love and sweetness (*adorada* (‘adored’) and *meiga* (‘sweet’)) stand out, but also words associated with sadness and insecurity (*pobre* (‘poor’), *chorosa* (‘tearful’), *ciumenta* (‘jealous’), *ofendida* (’offended’)), and fear (*espavorida* (‘terrified’)). On the other hand, bravery is the masculine highlight: *valente* (‘brave’) is, by far, the most frequent word and *atrevido* (‘cheeky/audacious’) is in the sixth place. Humility (*humilde* (‘humble’)) and anger (*furioso* (‘furious’)) rank second and third, respectively. Anxiety also appears: *desesperado* (‘desperate’) is the fourth most frequent emotional word for masculine characters.

Finally, masculine characters seem to be taken by surprise more often than feminine ones, frequently being *assombrado* (‘haunted’), *surpreso* (‘surprised’), and *maravilhado* (‘marveled’), which might be due to their role in narrative events.

‘Appearance’, although highly typical for feminine targets, varies relatively little in terms of the most frequently mentioned attributes: Beauty (*bonita*, *formosa*, *bela*, *linda*, Portuguese words for ‘beautiful’; *encantadora* (‘charming’)) or the lack of it (*feia* (‘ugly’)) are the most frequent features. In the masculine appearance axis, age and size, instead of beauty, are the most frequently mentioned attributes (*velho* (‘old’) and *jovem* (‘young’); *robusto* (‘robust’), *grande* (‘big’) and *baixo* (‘short’)).

On the typically masculine character axis, positive traits such as *grande* (‘great’), *simples* (‘simple’), *verdadeiro* (‘real’), *valente* (‘brave’), *livre* (‘free’), and *hábil* (‘skillful’) stand out. Other highly mentioned positive traits are *generoso* (‘generous’) and *habilidoso* (‘skillful’). Negative highlights are *mau* (‘bad’), *terrível* (terrible), and *rude* (‘rude’). For the feminine targets, the highlights are, in general, positive and associated with virtue: *virtuosa* (‘virtuous’), *santa* (‘holy’), and *inocente* (‘innocent’). Other typically feminine characterisation words are *meiga* (‘sweet’) and *dócil* (‘docil’), but we also see *fraca* (‘weak’), which contrasts with masculine strength.

4.3 Does the Gender of the Author Matter?

Do these findings vary according to the author’s gender? In our material, see Table 8, feminine authors use more appearance descriptions than masculine ones, as shown in Figure 13.

However, there is a huge difference in the size of the compared material: There are only 1.2 Mio. words written by women compared to almost 32 Mio. words written by men. In fact, this is an inescapable problem, given the reduced number of texts by women in our corpus: only 19 authors who wrote 33 works in prose. [16]

Even though the material is very unbalanced, we tried to discern any interesting trends in the works written by women in terms of whose appearance is described more – could it be that they would emphasise or concentrate more on the appearance of masculine

[16] Namely, ordered by decreasing number of words in the corpus: Júlia Lopes de Almeida, Virgínia de Castro e Almeida, Ana Plácido, Teresa Margarida da Silva e Orta, Maria Amália Vaz de Carvalho, Maria O’Neill, Maria Firmina dos Reis, Florbela Espanca, M.M.S.A. e Vasconcelos, Cláudia Campos, Maurícia C. de Figueiredo, Maria Luísa Marques da Silva, Matilde Isabel de Santana e Vasconcelos Moniz Bettencourt, Ana de Castro Osório, Alice Moderno, Maria Peregrina de Sousa, Paulina Filadélfia, Clarice Lispector, and Sônia Coutinho.

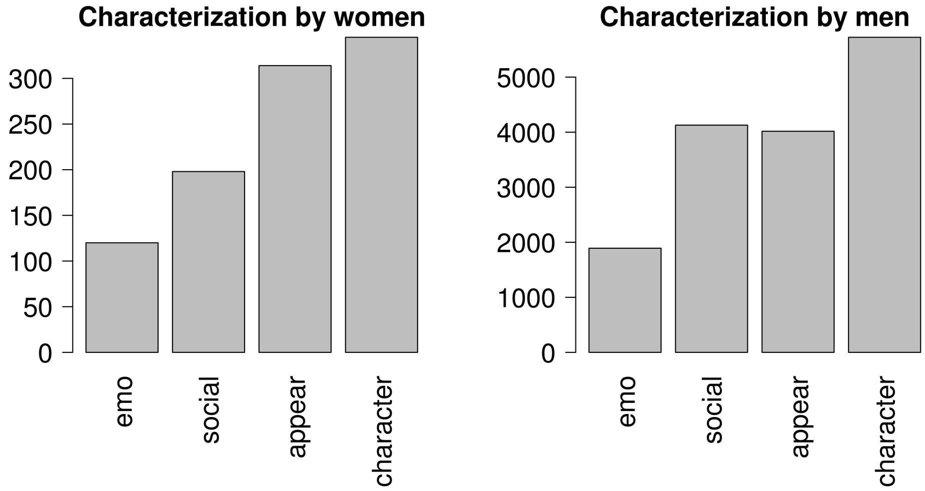

Figure 13: Characterisation by masculine and feminine authors. Note the different sizes in the y-axis.

	Total	Feminine author	Masculine author
Words	25,828,265	1,206,744	24,621,521
People	490,892	24,271	466,621
Characterised people	57,680	2,235	55,445
Social	8,968	355	8,613
Appearance	12,951	595	12,356
Emotion	8,704	533	8,171
Character	19,002	887	18,115

Table 8: Different depiction classes for masculine and feminine authors in novels, novellas, and short stories.

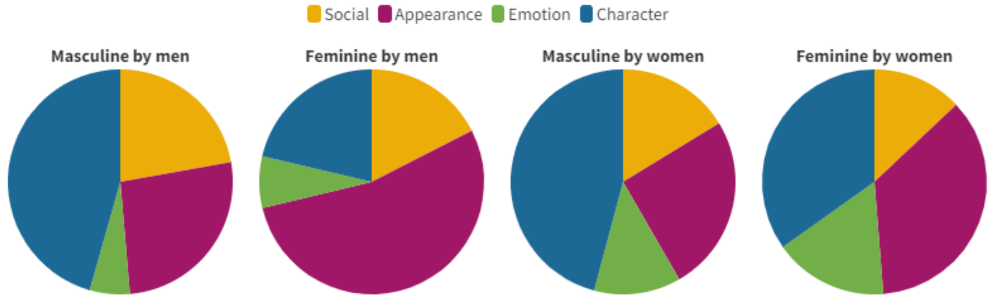

Figure 14: Preferred characterisation by masculine and feminine authors.

characters?

We get 265 appearance descriptions of feminine characters and 319 of masculine characters in 985 characterisations of feminine characters and 1,195 characterisations of masculine characters. In other words, 26.9% of feminine characterisations and 26.7% of masculine characterisations involve their appearance. But we acknowledge that the numbers are too small to be conclusive. In any case, it is conspicuous that both genders have roughly the same characterisation frequency in literature written by women.

Despite the imbalanced data, Figure 14 shows a preferential characterisation of both characters and writers in terms of gender. Below, we sketch some differences between human depiction in works written by men and women. The main difference is the increase of 'appearance' in masculine characterisation in works written by woman.

Beginning with feminine characters and focusing on women writers only, we found that *married* is no longer among the most frequent social depictions, but *widow* and *single* remain. Despite still being frequent, less space is devoted to beauty in works written by women. By contrast, age is more present: *young* and *old*. As for emotional characterisation, *happy* and *adorable* are the highlights, and none of the preferred emotional words relate to sadness. As for character, the highlights of feminine depiction words are *honest*, *infamous*, *crazy*, *refined*, and *dangerous*. In the social axis, masculine characters are mainly *married* and *noble*. Positive emotions are present for masculine characters as well (like *happy*/*pleased*, *enthusiastic*), but bravery (*brave*) has only one occurrence. Masculine 'appearance' follows the general trend, and masculine characters are mainly *kind* and *honourable*.

4.4 Differences between Brazil and Portugal

Are there differences between the two countries with regard to people's characterisation?

We compared the works from 1840 to the present day (Brazil became independent in 1822, and, as already mentioned, for the 1830 decade we only have one work by a Portuguese author).

We decided to compare only novels, novellas and short stories between the two countries because the non-fiction parts differ widely: While we have a large body of texts on history on the Portuguese side, we have almost only short essays in newspapers on the Brazilian side. The results are presented in Table 9.

	Total	Brazil	Portugal
People	486,575	209,283	277,292
Characterised people	46,704	19,642	27,062
Social	8,887	3,545	5,342
Appearance	12,877	6,199	6,678
Emotion	8,704	4,874	3,650
Character	18,782	7,649	11,133

Table 9: Different depiction classes in novels, novellas and short stories, in general, and per author nationality after 1840.

	Br total	Br fem.	Br masc.	Pt total	Pt fem.	Pt masc.
People	202,829	74,020	118,088	275,301	81,847	165,796
Characterised people	17,453	6,381	10,591	24,548	8,452	15,372
Social	3,545	1,216	2,217	5,342	1,753	3,434
Appearance	6,199	2,579	3,472	6,678	2,618	3,885
Emotion	3,474	1,444	1,949	5,230	2,206	2,925
Character	7,649	2,446	4,955	11,133	3,292	7,452

Table 10: Different depiction classes in novels, novellas, and short stories after 1840 per author nationality and per gender of the characterised.

We can see that the numbers of ‘character’ and ‘social’ characterisation are somewhat higher in Portuguese literature, while the other categories – especially emotion – are more pronounced in Brazilian literature. One may wonder whether this is due to a more socially rigid society in Portugal, or whether the cause lies in the historical novels (almost absent in the Brazilian material and quite frequent in the Portuguese material).

We also investigated whether the differences among genders are more obvious in the Brazilian material or different from the ones in the Portuguese material. For this, we created Table 10, where we can see that Brazilian literature has a higher proportion of mentions of feminine characters (36.5%) than the Portuguese (29.7%). This may again be due to the historical novels, but needs to be investigated further.

In Table 10, we see that the social status of male characters is more important in Portuguese literature.

If we now compare the distribution by country and by gender, presented in Figure 15, masculine characters seem to be similarly depicted, although in Portuguese-authored works there is a slightly more balanced distribution between appearance, social and emotion axes. In Brazilian-authored works, besides the emphasis on ‘appearance’, there is proportionally less use of the character axis, which leads to a smaller difference between characterisations by ‘appearance’ and by ‘character’. For feminine characters, there are relatively fewer mentions of their social status and emotional states in Brazilian-authored works.

4.5 Differences among Authors

In Table 11, we show the distribution of the types of characterisation for 12 canonical authors, six Brazilian and six Portuguese.

We can see that there are some differences among these authors. They agree in that none

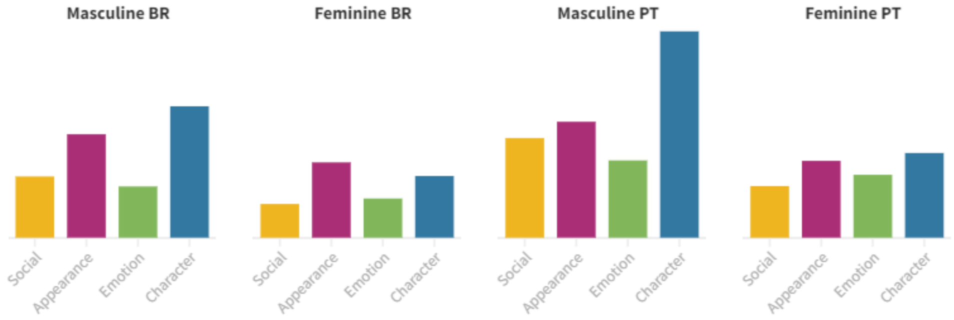

Figure 15: Characterisation by country.

Author	Country	nr	Total	Character	Social	Appearance	Emotion	mfreq
Camilo Castelo Branco	PT	42	4,045	1,781	938	845	481	pobre
Machado de Assis	BR	140	1,864	793	219	643	209	bom
Eça de Queirós	PT	16	2,487	1,019	420	923	125	bom
JM de Macedo	BR	7	1,325	411	232	515	167	velho
Aluísio Azevedo	BR	13	1,307	513	191	374	229	pobre
José d'Alencar	BR	15	887	331	154	370	32	velho
Coelho Neto	BR	17	966	369	81	440	76	velho
Humberto de Campos	BR	6	766	169	193	368	36	velho
Júlio Dinis	PT	9	1,038	430	127	302	179	pobre
Teófilo Braga	PT	4	419	144	82	112	81	pobre
Alexandre Herculano	PT	8	809	321	201	228	59	velho
Raul Brandão	PT	5	206	73	24	102	7	grande

Table 11: Different depiction classes per authors ordered by number of characterisations. "nr" shows the number of different fiction works by that author in *Literateca* and "mfreq" the most frequent characterising word.

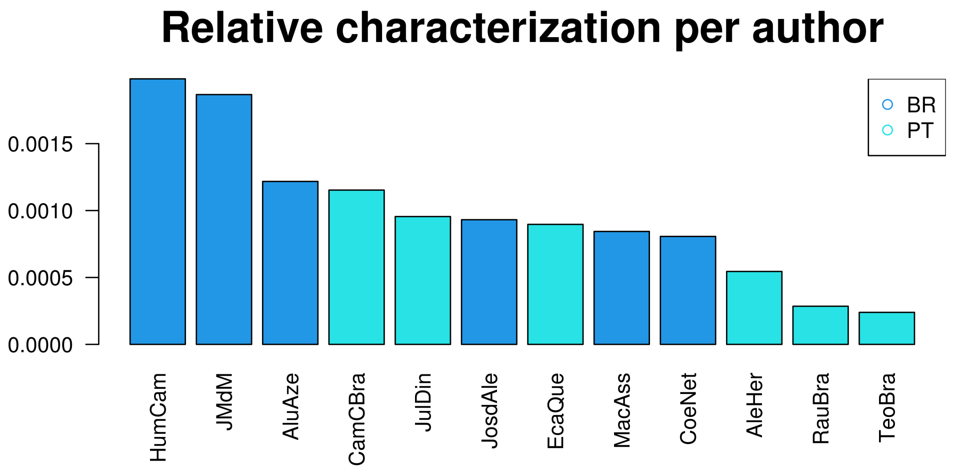

Figure 16: Characterisation by author. From left to right, Humberto de Campos, José Manuel de Macedo, Aluísio Azevedo, Camilo Castelo Branco, Júlio Dinis, José de Alencar, Eça de Queirós, Machado de Assis, Coelho Neto, Alexandre Herculano, Raul Brandão, and Teófilo Braga.

of them emphasises an explicitly emotional description, and several authors follow the "general" pattern in fiction: first 'character', then 'appearance', then 'social', and finally 'emotion': Machado de Assis, Eça de Queirós, Aluísio de Azevedo, José de Alencar, Júlio Dinis, Teófilo Braga, and Alexandre Herculano.

However, in José Manuel de Macedo, Coelho Neto and Raul Brandão 'appearance' is the most frequent characterisation and 'character' is the second most frequent.

As to the relative order of 'character' and 'social' characterisation, Humberto de Campos is the only one who reverts the "canonical" order, using more 'social' characterisations than those reflecting 'character', while Camilo Castelo Branco (incidentally the author with the highest number of works in *Literateca*) is the only one who describes more 'social' than 'appearance'.

In any case, there are also differences in the number of characterisations provided by each author: Figure 16 illustrates how much each author depicts, i.e. how many characterisations they use per number of words.

In Figure 17, we represent each author in a plane formed by internal and external characteristics.

4.6 The Influence of Literary School

For a subset of the works of *Literateca*, we have metadata about the literary school to which they belong, as described in Santos et al. (2020).

We selected all works marked as romantic in one group (11,850,395 words, 175 books) and those marked as realist or naturalistic (7,616,384 words, 121 different books) in another group[17] to see whether one could identify differences regarding people's depictions just based on this fourfold sub-classification, and also according to the gender of who gets characterised. The results are presented in Table 12 and in Figure 18.

[17] Note that the groups are not mutually exclusive: There are a few books classified as both romantic and realist, which correspond to the transition between the two schools.

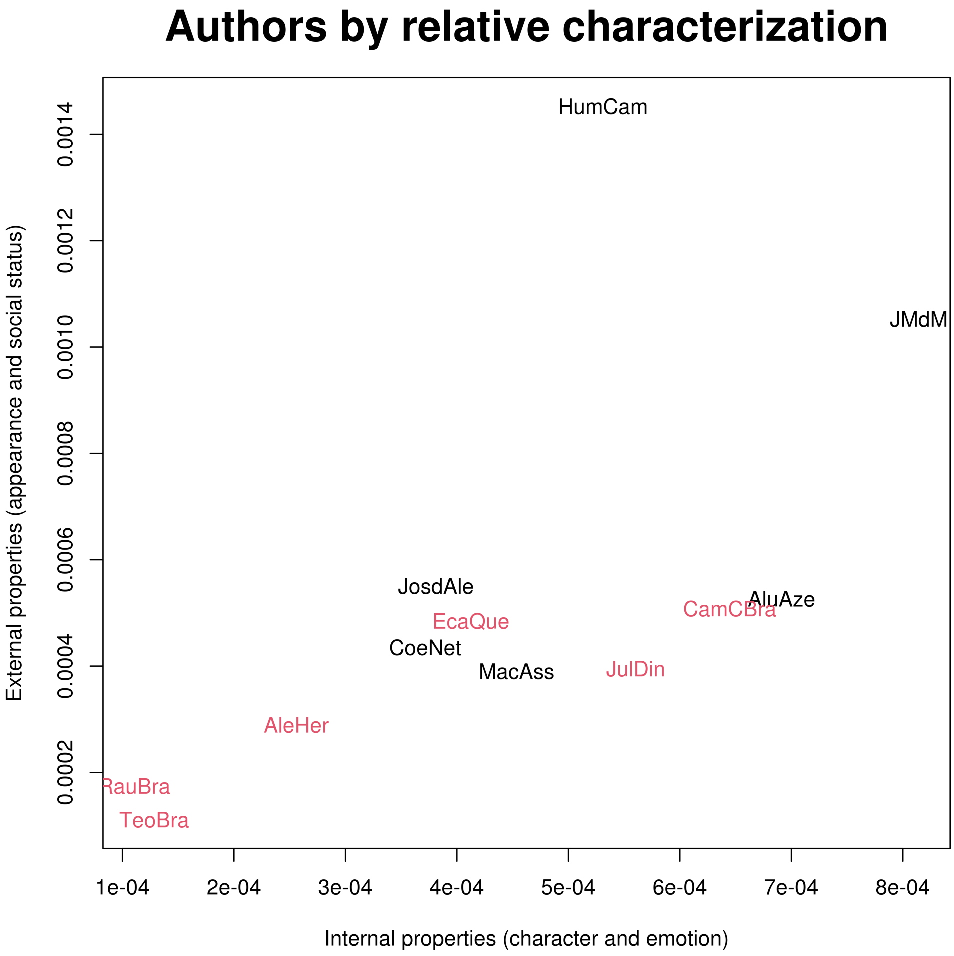

Figure 17: Characterisation by author in terms of type and relative weight of characterisation.

	Romantic	fem	masc	Realist	fem	masc
People	238,338	74,991	142,245	149,699	52,771	86,861
Characterised	22,733	8,140	14,041	13,834	5,187	8,244
Social	4,629	1,510	3,002	2,516	946	1,501
Appearance	5,573	2,279	3,179	3,944	1,678	2,147
Emotion	4,370	1,932	2,350	2,635	1,112	1,464
Character	9,389	2,899	6,237	5,649	1,819	3,650

Table 12: Different depiction classes in novels, novellas and short stories per literary school and per gender of the characterised.

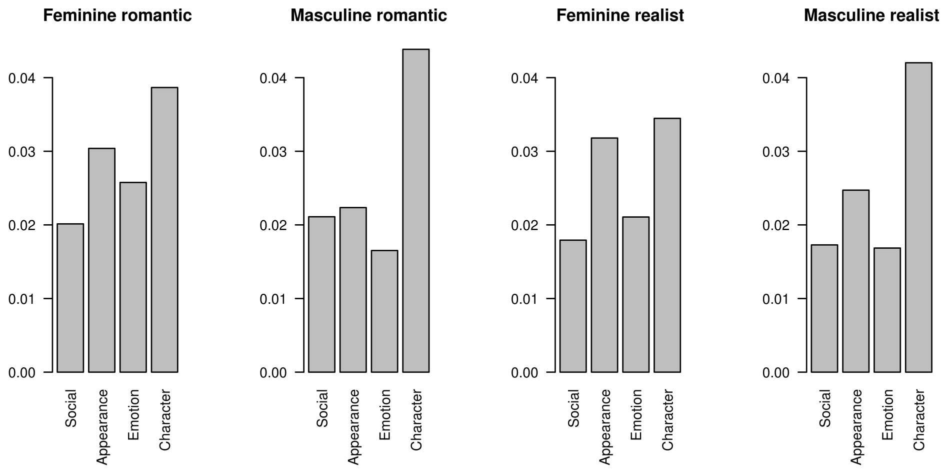

Figure 18: Characterisation per literary school and per gender.

The first interesting remark is that there are (relatively) more mentions of feminine characters in realist works than in romantic. However, 10.9% of the feminine occurrences are characterised in romantic books (and 9.9% of masculine occurrences), but only 9.8% in realist ones (compared to 9.5% for masculine).

We see that in romanticism, there are far more 'character' characterisations of masculine characters than in realism, where the relationship across all kinds of characterisations is stable across genres. In addition, realism describes the physical appearance of both genders, while romanticism prefers feminine appearance.

4.7 Going back to DIP

DIP has clearly demonstrated that there are fewer feminine characters in Lusophone literature. In this study, however, we see that those feminine characters are relatively more characterised, at least for 'appearance', than the masculine ones.

Ideally, and for the near future, we would like to connect the two studies/activities/forms of distantly looking at literature and provide, for each literary work, not only their description in terms of characters (as DIP does) but also how each character is characterised, using the present work and some form of anaphoric resolution of the non-proper name depictions and of those cases where human subjects (whether or not proper names) are omitted (Freitas and Souza (2021) found omitted subjects in 41% of clauses in Brazilian literature material).

We might therefore link types of characters with particular clusters of properties, like the beautiful rich woman and the poor honest lad and the evil old priest.

5. Concluding Remarks

In this paper, we offered some insights into human depiction based on distant reading literature in Portuguese. We can summarise our results as follows: Human depiction seems to obey the pattern 'character', 'social', 'appearance', and 'emotion' for masculine characters, and 'character' and 'appearance', 'social' and 'emotion' for feminine characters. If we consider only preferred depiction words, differences between feminine and masculine characters become more pronounced, and changing the lens – from distant to close reading – reveals that features associated with characters are related to their genders. The results also suggest an impact of the author's gender in the types of characterisation used, but the limited number of works written by women hinders a more definite conclusion.

We acknowledge that the material we used (works and words) is smaller than those used in other studies conducted under the umbrella of Digital Humanities. However, our findings show that an advantage of annotated data is the opportunity to see trends and patterns even in moderately sized collections. Furthermore, we stress that another intention of this work is to convince (the Portuguese-speaking community, mainly) to enlarge Portuguese-language literary collections with machine-readable texts.

In the near future, we would like to assess the precision of each rule used, and to correct the detected mistakes, as well as to widen the scope of characterisation. We are aware that human depiction is not restricted to the lexical-syntactic patterns we used, and to detect other ways in which the Portuguese language manifests characterisation is, therefore, a natural route to continue the investigation.

We are also aware that our study mainly reflects the vision of male authors of the nineteenth and early twentieth centuries. Therefore, it is by no means an unbiased description of gender. Other studies that we may undertake on this material will add an evaluative view: Which of these ways of depicting are positive, negative, or neutral? This is more straightforward for character and emotional words, but also possible for appearance and even social descriptions. We could also separate age from appearance and check what this dimension might bring.

In any case, all the material is open for inspection, from the lists of the characterising words to the patterns used, and the annotated works themselves, which allow interested researchers to repeat our searches and even refine them.

6. Data Availability

We make available in Zenodo:

- the list of characterising words, classified in five classes: `https://doi.org/10.5281/zenodo.7979566`;
- the patterns to find them in the corpus, together with the commands to create the tables and/or figures used in the paper: `https://doi.org/10.5281/zenodo.7979619`.

7. Acknowledgements

We thank Fundação Científica para a Computação Nacional (FCCN) of Portugal for hosting the corpora on which this study is conducted, and UNINETT Sigma2 – the National Infrastructure for High Performance Computing and Data Storage in Norway for the computational resources. We thank the reviewers for constructive criticism and the audience at the CCLS for comments and suggestions.

8. Author Contributions

Cláudia Freitas: Conceptualization, Writing – original draft, review & editing

Diana Santos: Conceptualization, Writing – original draft, review & editing

References

Argamon, Shlomo, Charles Cooney, Russell Horton, Mark Olsen, Sterling Stuart Stein, and Robert Voyer (2009). "Gender, Race, and Nationality in Black Drama, 1950-2006: Mining Differences in Language Use in Authors and their Characters". In: *Digital Humanities Quarterly* 3 (2). http://www.digitalhumanities.org/dhq/vol/3/2/000043/000043.html (visited on 01/17/2023).

Bick, Eckhard (2014). "PALAVRAS, a Constraint Grammar-based Parsing System for Portuguese". In: *Working with Portuguese Corpora*. Ed. by Tony Berber Sardinha and Thelma de Lurdes São Bento Ferreira. Bloomsbury Academic, 279–302.

Cao, Yang Trista and III Daumé Hal (2021). "Toward Gender-Inclusive Coreference Resolution: An Analysis of Gender and Bias Throughout the Machine Learning Lifecycle*". In: *Computational Linguistics* 47 (3), 615–661. 10.1162/coli_a_00413.

Cermáková, Anna and Michaela Mahlberg (2021). "The Representation of Mothers and the Gendered Social Structure of Nineteenth-Century Children's Literature". In: *English Text Construction* 14 (2), 119–149. 10.1075/etc.00044.cer.

— (2022). "Gendered Body Language in Children's Literature Over Time". In: *Language and Literature* 31 (1), 11–40. 10.1177/09639470211072154.

Freitas, Cláudia, Flávia Martins, and Liana Biar (2022). "Um 'olhar discursivo' sobre Predicação e Gênero: Aproximações Metodológicas entre Corpus e Discurso". In: *Texto Livre*. 10.35699/1983-3652.2022.36213.

Freitas, Cláudia and Elvis Souza (2021). "Sujeito oculto às claras: uma abordagem descritivo-computacional / Omitted subjects revealed: a quantitative-descriptive approach". In: *Revista de Estudos da Linguagem* 29 (2), 1033–1058. 10.17851/2237-2083.29.2.1033-1058.

Hoyle, Alexander Miserlis, Lawrence Wolf-Sonkin, Hanna Wallach, Isabelle Augenstein, and Ryan Cotterell (2019). "Unsupervised Discovery of Gendered Language through Latent-Variable Modeling". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1706–1716. 10.18653/v1/P19-1167.

Katsma, Holst (2018). *Loudness in the Novel*. https://litlab.stanford.edu/LiteraryLabPamphlet7.pdf (visited on 01/17/2023).

Larson, Brian (2017). "Gender as a Variable in Natural-Language Processing: Ethical Considerations". In: *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics, 1–11. 10.18653/v1/W17-1601.

Lucy, Li and David Bamman (2021). "Gender and Representation Bias in GPT-3 Generated Stories". In: *Proceedings of the Third Workshop on Narrative Understanding*. Association for Computational Linguistics, 48–55. 10.18653/v1/2021.nuse-1.5.

Mandell, Laura (2019). “Gender and Cultural Analytics: Finding of Making Stereotypes?” In: *Debates in the Digital Humanities*. Ed. by Matthew K. Gold and Lauren F. Klein. Manifold Scholarship, 3–26. 10.5749/j.ctvg251hk.4.

Moretti, Franco (2000). “The Slaughterhouse of Literature”. In: *Modern Language Quarterly* 61 (1). 10.1215/00267929-61-1-207.

— (2013). *Distant Reading*. Verso Books.

Moretti, Franco and Oleg Sobchuk (2019). “Hidden in Plain Sight: Data Visualization in the Humanities”. In: *New Left Review*, 86–115.

Rocha, Luísa, Cláudia Freitas, and Diana Santos (2019). “Preparação para Leitura Distante em Português: Diálogos entre PLN e Humanidades Digitais”. In: *Anais do TILic 2019*. 10400.26/31834.

Santos, Diana (2014). “Corpora at Linguateca: Vision and Roads Taken”. In: *Working with Portuguese Corpora*. Ed. by Tony Berber Sardinha and Thelma de Lurdes São Bento Ferreira. Bloomsbury Academic, 219–236. http://hdl.handle.net/10400.26/20539 (visited on 01/17/2023).

Santos, Diana and Cláudia Freitas (2019). “Estudando Personagens na Literatura Lusófona”. In: *Proceedings of the 12th Symposium in Information and Human Language Technology and Collocates Events (STIL)*, 48–52. https://comissoes.sbc.org.br/ce-pln/stil2019/proceedings-stil-2019-Final-Publicacao.pdf (visited on 01/17/2023).

Santos, Diana, Cláudia Freitas, and Eckhard Bick (2018). “OBras: a Fully Annotated and Partially Human-revised Corpus of Brazilian Literary Works in Public Domain”. In: *CorLex*. 10400.26/31830.

Santos, Diana, Cristina Mota, Emanoel Pires, Marcia Caetano Langfeldt, Rebeca Schumacher Fuão, and Roberto Willrich (2022a). *Introduction to DIP: Goal, Setup, Resources and Results*. https://www.linguateca.pt/aval_conjunta/dip/apr_encontro/DIPpresentation.pdf (visited on 01/17/2023).

— (2023). “DIP - Desafio de Identificação de Personagens: Objectivo, Organização, Recursos e Resultados”. In: *Linguamática* 15 (1), 3–30. 10.21814/lm.15.1.399.

Santos, Diana, Emanoel Pires, Cláudia Freitas, Rebeca Schumacher Fuão, and João Marques Lopes (2020). “Periodização Automática: Estudos Linguístico-Estatísticos de Literatura Lusófona”. In: *Linguamática* 12 (1), 81–95. 10.21814/lm.12.1.314.

Santos, Diana, Roberto Willrich, Marcia Langfeldt, Ricardo Gaiotto de Moraes, Cristina Mota, Emanoel Pires, Rebeca Schumacher, and Paulo Silva Pereira (2022b). “Identifying Literary Characters in Portuguese: Challenges of an International Shared Task”. In: *Proceedings of the 15th International Conference of Computational Processing of the Portuguese Language (PROPOR)*, 413–419. 10.1007/978-3-030-98305-5_39.

Schöch, Christof, Tomaz Erjavec, Roxana Patras, and Diana Santos (2021). “Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives”. In: *Modern Languages Open* 2021 (1), 1–19. 10.3828/mlo.v0i0.364.

Schöch, Christof, Evgeniia Fileva, and Julia Dudar (2022). “CLS INFRA D3.1 Baseline Methodological User Needs Analysis”. In: *Zenodo preprint*. 10.5281/zenodo.6389333.

Schulz, Daniel and Štepán Bahník (2019). “Gender Associations in the Twentieth-Century English-Language Literature”. In: *Journal of Research in Personality* 81, 88–97. 10.1016/j.jrp.2019.05.010.

Silva, Flávia Martins da Rosa Pereira da (2021). “Diferenciações de Gênero na Caracterização de Personagens: Uma Proposta Metodológica e Primeiros Resultados”.

MA thesis. PUC-Rio. https://www.maxwell.vrac.puc-rio.br/54130/54130.PDF (visited on 01/17/2023).

Smeets, Roel (2021). *Character Constellations: Representations of Social Groups in Present-Day Dutch Literary Fiction*. Leuven University Press. http://www.jstor.org/stable/j.ctv21wj5cb (visited on 12/17/2022).

Underwood, Ted (2019). *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press. 10.7208/9780226612973.

Underwood, Ted, David Bamman, and Sabrina Lee (2018). "The Transformation of Gender in English-Language Fiction". In: *Journal of Cultural Analytics* 3 (2). 10.22148/16.019.

Weingart, Scott and Jeana Jorgensen (2013). "Computational Analysis of the Body in European Fairy Tales". In: *Literary and Linguistic Computing* 28 (3), 404–416. 10.1093/llc/fqs015.