Gender depiction

Scale=MatchLowercase [] Ligatures=TeX, Scale=1

texgypagella [ Extension = .otf, UprightFont = *-regular, BoldFont = *-bold, ItalicFont = *-italic, BoldItalicFont = *-bolditalic]

texgypagella-math [ Extension = .otf, ]

FiraSans [ Extension = .otf, UprightFont = *-Regular, BoldFont = *-Bold, ItalicFont = *-Italic, BoldItalicFont = *-BoldItalic]

FiraCode [ Extension = .ttf, Path = ./fonts/firacode/, Contextuals = Alternate, UprightFont = *-Regular, BoldFont = *-Bold, ]

Ubuntu [ Extension = .ttf, Path = ./fonts/ubuntu/, UprightFont = *-R, BoldFont = *-B, ItalicFont = *-RI, BoldItalicFont = *-BI, ]

UbuntuMono [ Extension = .ttf, Path = ./fonts/ubuntu/, UprightFont = *-R, BoldFont = *-B, ItalicFont = *-RI, BoldItalicFont = *-BI, ]
Keywords: distant reading, annotation, Brazilian literature, Portuguese literature

Licenses: This article is licensed under: 

SUBMITTED

Gender Depiction in Portuguese
Distant reading Brazilian and Portuguese literature

Cláudia Freitas
Diana Santos

2. Department of Literature, University of Oslo, Oslo.

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 in 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 values and beliefs of those who describe. In addition to proper names, there are many other ways of human designation, such as the use of human general 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 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 insight. 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.

Underwood, Bamman, and Lee 2018, looking at English literature (104 thousand works, from 1703 to 2009), 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 as the volume of fiction written by women from 1850 to 1950 drops by a half.

Exploring the Black Drama collection, which contains plays written between 1950-2006, Argamon et al. 2009 reports poor results when trying to automatically distinguish the gender of the author and/or character. 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 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 million digitized books in English, analysed the lexical choice (adjectives and verbs) associated to 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 to preferences, “charming”, “fashionable” and “warm” were relatively feminine words, while “lazy” and “mean” were relatively masculine words. Men were described in decreasingly masculine terms throughout the beginning and end of the 20th 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 described more than masculine
characters with appearance-evaluative words, suggesting that men are associated with the mind and women with the body.

Cermáková and Mahlberg 2022 explores linguistic descriptions of gendered body language and compare 19th-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 sequence frequencies of 5 words which contain at least one body part noun and a marker of gender. They found fewer clusters for feminine characters in the 19th century. The contemporary data suggests, on the other hand, 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 unspecified 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 we add extra semantic annotation, which will be described below.

With Larson 2017, we recognize that using gender as a variable in Natural Language Processing is an ethical issue, and that we need to explicitly explain what “gender” means along this work. As Larson 2017 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. But we acknowledge 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 purpose to assert what gender is, but only how it was/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, Willrich, et al. 2022, to which we come back later.

Our point of departure is the work by Freitas, Martins, and Biar 2022 – and later extended in Silva 2021’s master thesis – who have suggested a fourfold classification for human characterisation. Human attributes were organised in social, appearance, character and emotional characteristics.

Using OBras, a corpus of Brazilian literature in the public domain Santos, Freitas, and Bick 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%), something which may be related to the fact that there were more masculine characters than feminine ones, roughly 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 reduces to married and rich.

- Emotional characterisations like feliz (‘happy’) were (almost) exclusively used for women.

We wanted to check whether these observations held 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, an evaluation contest for identifying literary characters and some information about them in Brazilian and Portuguese works Santos, Mota, et al. 2022; Santos, Mota, et al. 2023. By describing it and pointing out the differences, we throw some light on different ways to look at (roughly) the same data.

1. Although published in 2022, the work was conducted in 2018
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 relations with other characters. But the unit is the character. In addition, "literary character" in DIP does not include all people. In the present study, we try to look at all mentions of characterisation of people in the works, so our numbers are not per characters, but per mentions of people.

We will discuss and compare the findings about character gender in section 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, Erjavec, et al. 2021.

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\(^2\) 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 them, therefore contributing actively to the lack of information on lusophone materials and studies. For example, Schöch, Fileva, and Dudar 2022, page 4 state:

> As 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 the least because it shows the methodological problems it invites.

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

\(^2\) https://books.google.com/ngrams/
In order for these data to be shareable and studies replicable, we restrain our data (mostly\(^3\)) to books in the public domain. We are aware that many more electronic texts exist in electronic form, but by using them we would incur either on law infringement, or at least we would risk creating materials only for our own study, not shareable by 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 million tokens of (original) prose (excluding drama) from 1700 on.

A quantitative overview of the material is in Table 1.

<table>
<thead>
<tr>
<th>Literature</th>
<th>no. of tokens</th>
<th>no. works</th>
<th>no. authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>32,718,621</td>
<td>669</td>
<td>200</td>
</tr>
<tr>
<td>Portuguese</td>
<td>20,639,007</td>
<td>306</td>
<td>127</td>
</tr>
<tr>
<td>Brazilian</td>
<td>12,079,614</td>
<td>355</td>
<td>73</td>
</tr>
</tbody>
</table>

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

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

Literateca is the merge 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 which are not usually or necessary deemed literary, such as: newspapers chronics, letters, memoirs, and even scholarly works such as history books or ethnographic studies, and travel reports. For previous centuries, even sermons are included. However, these genres are only included for canonical writers.\(^4\)

3. Exceptions are excerpts of books existing in parallel corpora, or texts whose authors gave us permission to use.
4. By this we mean that established authors who belong to the Portuguese and Brazilian canons were fully digitized, that is, everything they published is available. This is in strong contrast with the works
• It includes drama, poetry and prose. 193

• Some of the works have updated orthography, others keep the original orthography. 194
  Given that there have been several norms of Portuguese spelling across the centuries, 195
  this means that there can be a variety of forms for the same word. 196

• While some authors have all their works included, others have only a few, or just 197
  one. Especially for non-canonical writers, there is no claim to completeness. 198

• Given that the works have been digitised by different bodies and with different 199
  tools and for different purposes, there is no claim to homogeneity: works can come 200
  from the first or the last paper version, they may keep their prefaces or not, they 201
  have different ways of describing chapters, etc. 202

• All works are marked with author, author gender, date of publication, variety of 203
  Portuguese, genre, and whether they are original or translated. Some texts are 204
  also classified by the literary school they belong to. 205

We tried to use as much as this material as we could, but we removed poetry and drama. 206
  Poetry is probably a natural choice to be removed, because of syntactic idiosyncrasies – 207
  and therefore a worse parser performance –, and because we believe that poetry has not 208
  so many mentions of fictional characters. We removed drama, also in prose, because it 209
  was heavily unbalanced, given that most of the plays were from Portugal. 210

As to prose, we started to use everything published since 1700. It is, anyway, important 211
  to recognise that we do not have a balanced corpus, and the lion’s part is fiction. We 212
  then selected different subsets for different research questions. 213

  • Just fiction, and just non-fiction, to see whether depiction was different across the 214
    fiction divide 215
  
  • Just works published after 1840, to be able to compare Brazilian and Portuguese 216
    authors 217

  • Just fiction published after 1840, to be able to compare Brazilian and Portuguese 218
    literature 219

See figures 2 and 3 for a bird’s eye view of the genre distributions in total and in fiction. 220

Only in Figure 3 do we include the variable author gender, since it is only in fiction that 221
  we have text written by women. 222

In Table 2 we give the numbers of words involved for the material published after 1840. 223

2.2. Gender attribution 224

We explore the influence of gender both in characters’ description and authorship. Masculine 225
  and feminine gender labels were manually ascribed to writers, for our corpus contains 226
  of non canonical authors, which may have had some of their (mainly) novels digitized in the context of 227
  other projects.
Figure 2: Genre in the full corpus. The unit is the work.

Figure 3: Genre in the fiction corpus. The unit is the work.
works written by canonical authors that have been 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 to the characters, gender labels were automatically assigned by PALAVRAS parser, and then manually revised by linguists (Rocha, Freitas, and Santos 2019; Silva 2021). The linguistic clues 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 – planet (‘planet’) is masculine, and trib (‘tribe’) is feminine. Therefore, to observe 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 of the 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 in two ways: (i) we enriched the lexicon of general human nouns, including names of professions as targets, and (ii) having extended the amount of works analysed to include works written by Portuguese authors, we broadened the lexicon of characterising words, based on prose of the eighteenth, nineteenth and twentieth centuries in Literateca. Along

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

<table>
<thead>
<tr>
<th></th>
<th>Fiction</th>
<th>Non fiction</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazil</td>
<td>10,547,327</td>
<td>1,532,287</td>
<td>12,079,614</td>
</tr>
<tr>
<td>Portugal</td>
<td>15,280,938</td>
<td>5,358,069</td>
<td>20,639,007</td>
</tr>
<tr>
<td>Total</td>
<td>25,828,265</td>
<td>6,890,356</td>
<td>32,718,621</td>
</tr>
</tbody>
</table>

5. Which, in turn, are an improvement of the patterns used in Freitas, Martins, and Biar 2022.
6. 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.
the process of data analysis, we were forced to discuss previous classification, which lead to precise classification guidelines and to 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. Some examples of what the patterns yielded follow (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 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. As to the assignment of the categories proper, follows their scope and the major decisions associated:

**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 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, 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 (‘snob’).
It is important to mention that each category works as a label, which in turn encodes 4 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 note that the words classified can often refer to non-human entities, as is the case of 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, such as example (8), since only they are retrieved by the patters 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 as characterising at all in the corpus. For example, initially we wanted to discard the words granítico ('made of granit') 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 granitic...]

(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 to the exclusion:

- 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 (‘polite’), etc.
- We did not classify relational adjectives, such as partidário (de...) (‘partison’), apologista (de...) (‘in favour of’), comparável (a ...) (‘comparable to’), emparelhado com (‘pairing with’), similhante a (‘similar to’), since a precise characterization would require a close reading of each sentence.

7. 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.

8. But note that educado and bem-educado, as words of size one, were included.
• We threw away misspellings, except for lack of diacritics. Our rationale being 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, underlying 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 for habits like madrugador (‘early riser’) and bébado (‘drunkard’ or ‘drunk’), which can be either due to biology or social bringing up. The word acanhado (shy), can be interpreted as a not-expansive person (character) or as someone fearful (emotion), and the same applies to impaciente (impatient), 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 any further studies on this matter.

After this classification, we ended up with a list of 4481 words which might be employed in depicting human beings, see Table 3. Due to the vagaries of the parser, we list the lemmas which can be verb infinitives for past participle forms, because we use the lemmas in our patterns.

<table>
<thead>
<tr>
<th>type</th>
<th>size</th>
</tr>
</thead>
<tbody>
<tr>
<td>social</td>
<td>1391</td>
</tr>
<tr>
<td>appearance</td>
<td>672</td>
</tr>
<tr>
<td>emotional</td>
<td>514</td>
</tr>
<tr>
<td>character</td>
<td>1578</td>
</tr>
<tr>
<td>other</td>
<td>326</td>
</tr>
<tr>
<td>total</td>
<td>4481</td>
</tr>
</tbody>
</table>

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

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 and computed how often and when the words were used to describe human beings.

9. That is, we considered missing accents something that could be present in the original paper edition, but not OCR mistakes.
11. 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.
We start by providing a picture of the distribution of human characters in time, in Figure 4, as well as how many depicting events we were able to identify, in Figure 5. A comment is in place: the decade of 1830 is a clear outlier, because it contains one short text only, of 19,334 words, a political pamphlet by Alexandre Herculano, in the whole decade. The same happens with 1950, which in the material is only represented by 4,777 words of Jorge Amado’s *Gabriela, Cravo e Canela*.

### 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 to feminine
characters.\textsuperscript{12} Figure 6 displays the overall distribution of characterisation words.

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Masc. characters</th>
<th>Fem. characters</th>
</tr>
</thead>
<tbody>
<tr>
<td>People</td>
<td>578,815</td>
<td>352,851</td>
<td>173,370</td>
</tr>
<tr>
<td>Characterised people</td>
<td>80,415</td>
<td>52,252</td>
<td>24,664</td>
</tr>
<tr>
<td>Social</td>
<td>11793</td>
<td>7813</td>
<td>3534</td>
</tr>
<tr>
<td>Appearance</td>
<td>15394</td>
<td>9099</td>
<td>5862</td>
</tr>
<tr>
<td>Emotion</td>
<td>9670</td>
<td>5562</td>
<td>3895</td>
</tr>
<tr>
<td>Character</td>
<td>23880</td>
<td>16542</td>
<td>6394</td>
</tr>
</tbody>
</table>

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

Figure 6: Distribution of characterisation words among the 4 classes, for all, masculine and feminine depictions.

The first observation is that there are way more masculine than 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 the by far 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’), belo (‘beautiful’), formoso (‘beautiful’), bonito (‘pretty’)).

Social characterisation comes third (most frequent words: rico (‘rich’), ilustre (‘illus- trious’), nobre (‘noble’), casado (‘married’), célèbre (‘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\%), something that corroborates the findings in previous studies, and to which we return in subsection 4.2.

\textsuperscript{12} 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, marking them as M/F. Also, remember that by “character” here we mean mentions to people, not distinct characters.

\textsuperscript{13} It may seem surprising at first to include age as appearance, but it is something that we visually assess.
4.1. Does textual genre matter?

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

On the one hand, we had left all 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 how you crossing Africa, compared with a narrative where you introduce fictional characters.

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

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Masculine</th>
<th>Feminine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td>25,828,265</td>
<td>291,403</td>
<td>159,216</td>
</tr>
<tr>
<td>People</td>
<td>490,892</td>
<td>5720</td>
<td>2979</td>
</tr>
<tr>
<td>Characterised people</td>
<td>47,450</td>
<td>30,036</td>
<td>16,620</td>
</tr>
<tr>
<td>Social</td>
<td>8968</td>
<td>7401</td>
<td>5226</td>
</tr>
<tr>
<td>Appearance</td>
<td>12,951</td>
<td>4922</td>
<td>3665</td>
</tr>
<tr>
<td>Emotion</td>
<td>8767</td>
<td>12587</td>
<td>5773</td>
</tr>
</tbody>
</table>

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.

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, that raises 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 trends similar to the ones we present 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 character.

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

Let us see now for the non-fiction part, 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: 16% and 18%, confirming that women are even less important in the public sphere.

Now 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 women that are mentioned have a large proportion of appearance descriptions, even more in non-fiction than in fiction.

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

4.2. Differences when describing masculine and feminine characters

The previous figures have shown that appearance is more frequent when describing feminine characters. This can also be appreciated in the barplot of figure 9.

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 bunch of depictive words. In order to be evaluated as ‘preferred’, a word must (i) be used for masculine targets at least for 80% of the occurrences, or for feminine targets more than 60% of the occurrences; (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 they could be included in the preference count.
The new data are 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 in 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) and appearance, the second most frequent characterisation (of both masculine and feminine characters), drops to the third position when associated with masculine characters, and up to the first position, when associated with feminine characters.

Figure 10: Preferred characterisation per gender

The appearance axis is the second most common for both genders, but figures 11 and 12, complementary to figure 10, provide a few details that enrich the analyses. 14.

14. In figures 11 and 12, words such as beautiful₁ and pretty₂ 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.
As noted in previous studies, typically feminine social characterisations relate to the familiar environment (mãe ('mother'), prima ('cousin')), but mentions to marital status are the highlight (casada ('married') and viúva ('widow') are the most frequent words, but adultera ('adulteress') is frequent as well). Marital status, in turn, is absent as typical masculine social characterisation. These are related to (positive) social recognition such as ilustre ('illustrious'), célebre ('famous'), famoso (another word for 'famous') and poderoso ('powerful').

On the feminine emotional axis, words associated with love and sweetness (adored and sweet) stand out, but also words associated with sadness and insecurity (poor, tearful, jealous, offended) and fear (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 6th. Anxiety also appears (desesperado ('desperate') is the third most frequent).

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

Appearance, although highly typical for feminine targets, varies relatively little as to the most mentioned attributes: beauty (bonita; formosa, bela, linda, Portuguese words for beautiful) or the lack of it (feia ('ugly') is the most frequent feature. In the masculine appearance axis, age and size, instead of beauty, are the most mentioned attributes: velho ('old') and jovem ('young'); robust, big and short.

In the character axis, typically masculine, stands out grande ('great'), honrado ('hon-
ourable'). Other highly mentioned positive traits are *generoso* ('generous'), *habilidoso* ('skillful'), *real, sério* ('serious') and *leal* ('loyal'). For the feminine targets, the highlights are *virtuosa* ('virtuous'), *inocente* ('innocent') and *meiga* ('sweet').

### 4.3. Does the gender of the author matter?

Do these findings change 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.

![Characterisation by feminine and masculine authors](image)

**Figure 13:** Characterisation by masculine and feminine authors. E PRECISO REFAZER
However, there is a huge difference in the size of the material compared: there are only 1.2 million words written by women compared to almost 32 million words by men. In fact, this is an inescapable problem, given the reduced number of writings by women in our corpus: only 19 authors\(^{15}\) who wrote 33 works in prose.

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Feminine author</th>
<th>Masculine author</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td>25,828,265</td>
<td>1,206,744</td>
<td>24,621,521</td>
</tr>
<tr>
<td>People</td>
<td>490,892</td>
<td>24,271</td>
<td>466,621</td>
</tr>
<tr>
<td>Characterised people</td>
<td>57,680</td>
<td>2235</td>
<td>55,445</td>
</tr>
<tr>
<td>Social</td>
<td>8968</td>
<td>355</td>
<td>8613</td>
</tr>
<tr>
<td>Appearance</td>
<td>12951</td>
<td>595</td>
<td>12356</td>
</tr>
<tr>
<td>Emotion</td>
<td>8704</td>
<td>533</td>
<td>8171</td>
</tr>
<tr>
<td>Character</td>
<td>19002</td>
<td>887</td>
<td>18115</td>
</tr>
</tbody>
</table>

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

Even though the material is heavily unbalanced, we tried to discern any interesting trend in works written by women as far as whose appearance was more described – could it be that they would emphasise or concentrate more on the appearance of masculine characters?

We get 265 appearance descriptions of feminine characters, and 319 of masculine characters, in 985 characterisations of feminine characters and 1195 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 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 preferential characterisation regarding gender of both characters and writers. In what follows 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. On the other hand, age is more present: *young* and *old*. As to emotional characterisation, *happy* and *adorable* are the highlights, and none of the preferred emotional words relate to sadness. As to character, the highlights of feminine depiction words are *honest, infamous, crazy, refined* and *dangerous*. In the social axe, masculine characters are mainly *married* and *noble*. Positive emotions are present for masculine characters as well (*happy/pleased, enthusiastic*), but bravery (*brave*) has only

\(^{15}\) Júlia Lopes de Almeida, Virginia 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 Filadellia, Clarice Lispector and Sônia Coutinho, by decreasing number of words in the corpus.
one occurrence. Masculine appearance follows the general trend, and masculine character are mainly kind and honourable.

![Figure 14: Preferred characterisation by masculine and feminine authors.](image)

**4.4. Difference between Brazil and Portugal**

Are there differences between the two countries as regards people’s characterisation?

We compared the works from 1840 to the present (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 in the Portuguese side, we have almost only short essays in newspapers on the Brazilian side. The results are presented in Table 9.

<table>
<thead>
<tr>
<th>Total</th>
<th>Brazil</th>
<th>Portugal</th>
</tr>
</thead>
<tbody>
<tr>
<td>People</td>
<td>486,575</td>
<td>209,283</td>
</tr>
<tr>
<td>Characterised people</td>
<td>46,704</td>
<td>19,642</td>
</tr>
<tr>
<td>Social</td>
<td>8887</td>
<td>3545</td>
</tr>
<tr>
<td>Appearance</td>
<td>12877</td>
<td>6199</td>
</tr>
<tr>
<td>Emotion</td>
<td>8704</td>
<td>4874</td>
</tr>
<tr>
<td>Character</td>
<td>18782</td>
<td>7649</td>
</tr>
</tbody>
</table>

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

We see that 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 if this is due to a more socially rigid society in Portugal, or whether the cause lies with 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 will have to be investigated closer.
Here we see that the social status of male characters is more important in Portuguese literature.

If we now compare the distribution by country by gender, presented in Table 10, masculine characters seem to be similarly depicted. But for feminine characters, there are significantly relatively less mentions of their social status and more mentions of their appearance in Brazilian authored works.

Figure 15: Characterisation by country

### 4.5. Differences among authors

In Table 11, we show the distribution of the kinds of characterisation for 12 canonical authors, 6 Brazilian and 6 Portuguese.

We see there are some differences among these authors. They agree in that none emphasises an explicitly emotional description, and several authors follow the "general" pattern in fiction: first character, then appearance, then social, and last emotion: Machado de Assis, Eça de Queirós, Aluísio de Azevedo, José de Alencar, Júlio Dinis, Teófilo Braga and Alexandre Herculano.

But 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

---

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>People</td>
<td>202,829</td>
<td>74,020</td>
<td>118,088</td>
<td>275,301</td>
<td>81,847</td>
<td>165,796</td>
</tr>
<tr>
<td>Characterised</td>
<td>17453</td>
<td>6381</td>
<td>10591</td>
<td>24548</td>
<td>8452</td>
<td>15372</td>
</tr>
<tr>
<td>Social</td>
<td>3545</td>
<td>1216</td>
<td>2217</td>
<td>5342</td>
<td>1753</td>
<td>3434</td>
</tr>
<tr>
<td>Appearance</td>
<td>6199</td>
<td>2579</td>
<td>3472</td>
<td>6678</td>
<td>2618</td>
<td>3885</td>
</tr>
<tr>
<td>Emotion</td>
<td>3474</td>
<td>1444</td>
<td>1949</td>
<td>5230</td>
<td>2206</td>
<td>2925</td>
</tr>
<tr>
<td>Character</td>
<td>7649</td>
<td>2446</td>
<td>4955</td>
<td>11133</td>
<td>3292</td>
<td>7452</td>
</tr>
</tbody>
</table>

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

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

Figure 16: Characterisation by author.

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 they belong, as has been described in Santos, Pires, 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, \footnote{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.} to see whether one could identify differences as to 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.

The first interesting remark is that there are (relatively) more mentions of feminine characters in realist works than in romanticism. However 10.9% of the feminine occurrences are characterized 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 characterizations is stable across genres. In addition, realism describes physical appearance of both genders, while romanticism prefers feminine appearance.

\textbf{Figure 17:} Characterisation by author as far as kind and relative weight of characterisation.

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

<table>
<thead>
<tr>
<th></th>
<th>Romantic</th>
<th>fem</th>
<th>masc</th>
<th>Realist</th>
<th>fem</th>
<th>masc</th>
</tr>
</thead>
<tbody>
<tr>
<td>People</td>
<td>238,338</td>
<td>74,991</td>
<td>142,245</td>
<td>149,699</td>
<td>52,771</td>
<td>86,861</td>
</tr>
<tr>
<td>Characterised</td>
<td>22,733</td>
<td>8140</td>
<td>14041</td>
<td>13834</td>
<td>5187</td>
<td>8244</td>
</tr>
<tr>
<td>Social</td>
<td>4629</td>
<td>1510</td>
<td>3002</td>
<td>2516</td>
<td>946</td>
<td>1501</td>
</tr>
<tr>
<td>Appearance</td>
<td>5573</td>
<td>2279</td>
<td>3179</td>
<td>3944</td>
<td>1678</td>
<td>2147</td>
</tr>
<tr>
<td>Emotion</td>
<td>4370</td>
<td>1932</td>
<td>2350</td>
<td>2635</td>
<td>1112</td>
<td>1464</td>
</tr>
<tr>
<td>Character</td>
<td>9389</td>
<td>2899</td>
<td>6237</td>
<td>5649</td>
<td>1819</td>
<td>3650</td>
</tr>
</tbody>
</table>

Figure 18: Characterisation per literary school and per gender

4.7. Going back to DIP

DIP has clearly demonstrated that there are fewer feminine characters in lusophone literature.

But in this study 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 kinds 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 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 moderate-sized collections. On the other hand, we stress that another intention with this work is to convince (the Portuguese-speaking community, mainly) to enlarge literary Portuguese-language 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 Portuguese language manifests characterisation is, therefore, a natural route to continue the investigation.

We are also aware that our study reflects mainly the vision of male authors of nineteenth and early twentieth century, and that therefore it is by no means an unbiased description of gender.

Other studies that we may still do on this material is to add an evaluation view: of these ways of depicting, which ones 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 may bring.

Anyway, all 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 redo our searches and even refine them.
6. Software availability

Not relevant

7. Acknowledgements

Funding, Funding and thanks!

8. Author contributions

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

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

References


