Character Constellations

Smeets, Roel

Published by Leuven University Press


For additional information about this book
https://muse.jhu.edu/book/98048

For content related to this chapter
https://muse.jhu.edu/related_content?type=book&id=3191737
2.1 INTRODUCTION: DESCRIPTIVE STATISTICS

How are social groups represented in present-day Dutch literary fiction? In order to provide an answer to this question, the subsequent third, fourth, and fifth chapters – on centrality, community, and conflict – each break down one aspect of this representation. As each of these chapters is based on the demographic metadata and relational information on 2,137 characters as annotated in several periods of data collection (see section 1.4.2 of the introductory chapter), a description of this dataset is first required. In this chapter, some basic descriptive statistics on the characters populating the 170 novels are presented. A closer look at these descriptives is not only a convenient introduction to the data which the subsequent analyses are based on but also provides a first, general sense of how characters of a certain gender, descent, education, and age are depicted in the corpus.

Information on the authors of the 170 novels in the corpus is provided first. Then, a broad overview on the demographics of the 2,137 characters is reported, and some basic statistical tests are conducted to determine whether the occurrences of characters from a certain gender, descent, education, and age deviate significantly from their hypothesized occurrences. After this, tests are performed to determine whether or not gender, descent, education, and age of characters are statistically (in)dependent of one another. Following the demographic overview and the tests of independence, relational information
is provided on how the roles of family, colleague, friend, lover, and enemy are distributed among these characters. Finally, this chapter concludes with a reflection on the reported descriptive statistics.

2.2 INFORMATION ON THE AUTHORS

The 170 novels were all written in the Dutch language. Of these books, 5 novels were written in collaboration, which leads to a total of 175 authors in the database. The gender divide among the authors is almost 70:30; 122 authors are male (69.7%) and 53 female (30.3%). The majority of the authors were born in the Netherlands (76.0%) or Flanders (16.6%). A small portion of 7 authors originated from a non-Western country (4.0%). In terms of education, it proved to be impossible to determine educational level for 14.8% of the male and 15.1% of the female authors. For those whose education could be retrieved, both male and female authors are higher educated (96.2% male, 100% female). Many authors live in Amsterdam (28%), a smaller number lives outside one of the large Dutch cities in the Randstad (17.1%), followed by a share of authors living in a large city in the Randstad other than Amsterdam (10.3%) or one of the large cities of Belgium (8%).

2.3 DEMOGRAPHIC METADATA ON THE CHARACTERS

In this book, the 170 novels in the corpus are used as a sample population of present-day Dutch literary fiction, and the characters in those novels are subsequently considered as a sample of the population of fictional characters in present-day Dutch language fiction. This population consists of 2,137 characters of which 59.80% is male and 40.10% is female; for only two characters the gender could not be determined (0.09%). Which gender distribution would we expect? Based on the hypothesis that authors tend to write more about characters of their own gender, the overrepresentation of male authors (69.7%) in the corpus might suggest an overrepresentation of male characters as well. An alternative null hypothesis is that the gender distribution among characters reflects the gender distribution in the society the books were published in. This hypothesis,
Data

however, assumes that the fictional population of characters is a reflection of actual demographics, which is a thought-provoking but theoretically problematic assumption. Leaving aside assumptions on the effect of author gender on character gender as well as effects of real-world population demographics, there is no reason to assume that there would not be an equal gender distribution between characters. In order to test if the null hypothesis of an equal gender distribution among characters holds, Pearson’s chi-squared goodness of fit test was calculated comparing the occurrence of male and female characters with the hypothesized occurrence of a 50–50 gender distribution. Significant deviation from the hypothesized values was found \( \chi^2 (1) = 82.030, p < 0.001 \), which means that the gender ratio of, roughly, 60:40 is a statistically significant difference and is thus very unlikely to be due to chance. Male characters are, in other words, significantly more present in present-day Dutch literary fiction.

Figure 1 shows that the great majority of the characters originate from or live in the Netherlands (52.76% and 55.79% respectively), followed by characters originating from or living in Belgium (8.95% and 9.61% respectively), other countries in Europe (10.26% and 10.31% respectively), or non-Western and non-Middle Eastern countries (8.39% and 10.13% respectively; categorized as ‘Other’). It is noteworthy that a relatively small portion of countries of descent and residence is unknown. Apparently, these are character features that are made explicit relatively often throughout the novels.

![Figure 1. Character distributions for country of descent and country of residence (N = 2,137).](image-url)
Following the same line of reasoning as for character gender, a reasonable null hypothesis is that distributions among places of origin and places of residence are equal for all categories. Of course, it can be argued that chances are high that Dutch language novels feature Dutch or Belgian characters because of the Dutch or Belgian background of their authors. An equally compelling argument, however, is that literary fiction is not bound to real world demographics, and that we might expect characters of a wide range of places of origin and residence in the Libris corpus. Following that argument, Pearson’s chi-squared goodness of fit test was calculated comparing the occurrence of characters with a Dutch, Belgian, European, Western, Middle Eastern, or ‘Other’ country of descent with the hypothesized occurrence of an equal distribution among those categories. Significant deviation from the hypothesized values was found ($\chi^2 (5) = 2599.865, p < 0.001$). The same test, with the same hypothesized occurrence, was carried out for country of residence. This test also demonstrated a significant deviation from the hypothesized values was found ($\chi^2 (5) = 2830.463, p < 0.001$). These tests indicate that the unequal distribution among countries of descent and residence is not due to chance, but points at a statistical difference. Given the majority of characters born or living in the Netherlands (52.76% and 55.79% respectively), these tests show that characters in present-day Dutch novels significantly more often originate from or live in the Netherlands than that they originate from, or live in, other countries.

For level of education the portion of unknown takes up 40.39% (see Figure 2). This suggests that education is a relatively less articulated or significant aspect of characters. This also applies to age (see Figure 3): for the largest part of the characters this is an unknown demographic feature (37.02%).

![Figure 2. Character distributions for level of education, divided by gender (N = 2,137).](image-url)
Again, there is no reason to assume that education or age would be unequally distributed among the characters. A chi-squared goodness of fit test was calculated comparing the occurrence of characters in age categories <25, 26–35, 36–45, 46–55, 56–64, 65+ with the hypothesized occurrence of an equal distribution among those categories. A significant deviation from the hypothesized values was found ($\chi^2 (6) = 1218.344, p < 0.001$). For education, the same chi-squared goodness of fit test was calculated comparing the occurrence of characters with a high and a low level of education with the hypothesized occurrence of an equal distribution among those categories. Again, a significant deviation from the hypothesized values was found ($\chi^2 (2) = 247.890, p < 0.001$). The results of these two tests indicate that the earlier found patterns of young and higher educated characters being overrepresented are thus significant.

2.4 (IN)DEPENDENCE OF VARIABLES

In order to determine whether or not two variables are (in)dependent on one another (e.g., gender on age or descent on education), Pearson’s chi-squared tests of independence were conducted for a range of combinations of two variables. The outcomes of these tests give a general insight in the intersections between demographic features and identity categories. For each of the four targeted categories in the analyses presented in the subsequent chapters (gender, country
of descent, education, age), chi-squared tests of independence were computed for a range of combinations of the variables gender, education, age, and country of descent.\footnote{For gender, associations between the variables education and age were tested. It would be interesting to see if gender is dependent on educational status, as well as to see if gender is dependent on youthfulness or maturity. There appears to be a significant association between character gender and character age ($\chi^2 (6) = 64.724, p < 0.001$), as well as between character gender and character education ($\chi^2 (2) = 60.579, p < 0.001$). This means that a character's gender and a character's age, as well as a character's gender and a character's education are not independent from one another but are significantly associated. These findings, however, only pertain to the general dependence of gender on respectively age and education, but do not reveal whether or not, for instance, female characters are more often higher educated and older than male characters. Below, the statistical significance of such differences is reported.

Looking at education and age from this binary gender perspective, some basic trends stand out. In general, the characters are higher educated (42.08%). Interestingly, level of education is more unknown for female than for male characters (see Figure 2), which is also a statistically significant difference (see Appendix A). Apparently, education is less mentioned or made less often explicit for female characters than for male characters. The opposite holds true for age (see Figure 3). In general, most characters are in age group <25 (23.01%) and the smallest portion of characters is represented in age group 56–64 (4.26%). More specifically, age is considerably more unknown for male characters (41.60%) than for female characters (29.94%), which appears to be a statistically significant difference (see Appendix A). These statistically significant differences demonstrate that age is more often mentioned for female than for male characters. Furthermore, female characters are on average younger than male characters, which is best visible in the overrepresentation of female characters in age categories <25, 26–35 and 36–45 (all differences are statistically significant, see Appendix A).

Besides interdependencies of gender on the one hand and education and age on the other, it is insightful to see how education and descent are statistically associated with one another. One hypothesis is that class, as indicated by education, is dependent on place of birth. A significant association between character education and character country of descent was found ($\chi^2 (10) = 99.562, p < 0.001$), as well as between character education and character city of descent ($\chi^2 (22) = 81.039, p < 0.001$). A character’s education and a character’s
country of descent, as well as a character’s education and a character’s city of descent, are thus not independent of one another but are significantly associated.

Breaking down these associations between education and descent, Figure 4 shows that 45.3% of Dutch characters is higher educated. This is a statistically significant difference from the 29.8% of Belgian and 29% of Middle Eastern and ‘Other’-descent characters who are higher educated (see Appendix A).

![Figure 4. Character distributions for education, divided by country of descent (N =2,137).](image)

The differences between European characters with a higher education (46.6%), Western characters with a higher education (52.6%), and Dutch characters is not statistically significant, indicating that for those characters a higher education is an equally mentioned character feature. A similar pattern holds for lower educated Dutch (13.3%), European (17.8%), and Western (7%) characters; those differences are again not statistically significant. Similar to higher education, the amount of lower educated Dutch characters (13.3%) differs significantly from lower educated Belgian (22.5%), Middle Eastern (25%), and ‘Other’ (39.1%) characters. Apparently, characters with Belgian, Middle Eastern, and ‘Other’ roots are significantly more likely to be lower educated than characters with Dutch roots.

The demographic landscape thus far presents an image that is complementary to the findings from earlier, preliminary research on the same corpus (Van der Deijl et al., 2016), which was based on a less extensive dataset. The reported
descriptive statistics on the present dataset do not dispute the general trends which were reported earlier: present-day Dutch literary fiction is predominantly populated by male, higher educated, Dutch characters (39). In this preliminary study, no Pearson chi-squared tests were conducted. These current tests statistically confirm the patterns that were observed earlier on a larger scale.

2.5 RELATIONAL INFORMATION

The added value of the present dataset not only lies in the larger number of characters, but also in the appended relational labels between characters. Appendix B contains the total distributions among the relational roles colleague, friend, lover, enemy, family. This general overview shows that family is the most prevailing relational role (42.39%), followed by colleague (25.49%), friend (16.56%), lover (9.7%), and enemy (5.86%).

Breaking these relational data down by gender, a remarkable trend stands out. The gender distributions among characters sharing the above-mentioned relational roles are represented in Table 1. As opposed to male-male relations (38.57%) and opposite-sex relations (43.27%), female-female relations take up a minor part in these types of character relations (17.77%).

<table>
<thead>
<tr>
<th>Relation type</th>
<th>Frequency distribution</th>
<th>By male authors</th>
<th>By female authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>male-male</td>
<td>38.57%</td>
<td>47.54%</td>
<td>19.72%</td>
</tr>
<tr>
<td>male-female / female-male</td>
<td>43.27%</td>
<td>41.60%</td>
<td>46.00%</td>
</tr>
<tr>
<td>female-female</td>
<td>17.77%</td>
<td>10.49%</td>
<td>33.79%</td>
</tr>
</tbody>
</table>

Table 1. Distribution of gender-gender character relations (N = 2,137), divided by gender author (N= 175).

These results bring to mind the famous Bechdel test for testing female presence in artworks, which is used to validate whether or not a story features a scene in which two women speak about something other than a man. The namesake of the test, cartoonist Alison Bechdel,15 was inspired by the feminist writings of Virginia Woolf, whose quote from *A Room of One’s Own* first sparked the idea on which the Bechdel-test is based:
All these relationships between women, I thought, rapidly recalling the splendid gallery of fictitious women, are too simple. [...] And I tried to remember any case in the course of my reading where two women are represented as friends. They are now and then mothers and daughters. But almost without exception they are shown in their relation to men. It was strange to think that all the great women of fiction were, until Jane Austen’s day, not only seen by the other sex, but seen only in relation to the other sex. And how small a part of a woman’s life is that [...] (Woolf, 1929, chapter 5; my emphasis)

Woolf signals that female characters in literary fiction are often primarily defined in their relation to male characters. Although Woolf had no dataset to back up her statements, the notion that the importance of women in fiction is relative to the male perspective has become common sense, of which the Bechdel test is a clear illustration. For the present dataset, this notion is supported by the relational data between male and female characters. Although Woolf’s statement that female characters are ‘almost without exception [...] shown in their relation to men’ does not hold completely for this dataset, it is the case that relations between female characters are underrepresented in relation to male-male and opposite-sex relations.16

Filtered out by author gender, Table 1 also shows that male authors tend to write more about male-male (47.54%) or opposite sex relations (41.60%), and that female authors write more often about female-female (33.79%) or opposite sex relations (46.00%). Both male and female authors write sparsely about relations between two characters of the other gender.

In the subsequent chapters, the backgrounds of authors will not be a main focus in the analyses. In line with this book’s text-centric focus,17 the presented models will primarily target representations of social groups on the level of the text. The findings presented in Table 1, however, do suggest that at least author gender has an effect on these representations.

2.6 INTERPRETATION OF DESCRIPTIVE STATISTICS

The reported descriptive statistics provide a first, general image of the representation of social groups in present-day Dutch literary fiction. Anticipating the following chapter on centrality, the frequency distributions of the demographic categories
are, in a broad sense, indicative of the dominance of certain character types over others. The 60:40 gender distribution indicates that male characters are more present, more visible, and thus possibly more central, than female characters. Another indication is given by the distributions within the categories of country of descent and country of residence: there is a significant overpopulation of Dutch characters as opposed to characters from other countries of descent and residence. These findings, however, only relate to frequency of occurrence. Male and Dutch characters are relatively central in terms of their frequency of occurrence in the corpus as a whole, but this does obviously not shed light on how male and Dutch characters in individual novels are represented.

The frequency distributions also suggest patterns that pertain to stereotypes or biases related to gender, class, and cultural background. This is particularly demonstrated by the amount of unknown or missing data with regard to education and age. From a gender perspective, these missing data invoke the bias that age is a more important demographic category for women than for men, which is also strengthened by the relatively large number of young female characters. In a similar vein, the discrepancy in unknown age between male and female characters relates to the bias that class, as expressed by educational status, is a more relevant demographic category for men than for women. Furthermore, the overrepresentation of higher educated Dutch, European, and Western characters as opposed to the underrepresentation of lower educated Middle Eastern and ‘Other’ characters reproduces existing class hierarchies between people of ‘Western’ and ‘non-Western’ descent. Belgian characters form an interesting exception to this pattern: they are in the same position as Middle Eastern and ‘Other’ character with regard to educational status but are clearly not part of the ‘non-Western’ category.

Descriptive statistics regarding the relational data indicate that family is the most central relational role which characters perform. This finding alludes to the commonplace of the family being the cornerstone of a society, which is described by former American president Lyndon B. Johnson in the following terms:

The family is the cornerstone of our society. More than any other force it shapes the attitude, the hopes, the ambitions, and the values of the child. And when the family collapses it is the children that are usually damaged. When it happens on a massive scale the community itself is crippled. (Johnson, 1965)
In this view, the social structure of the family is crucial for a society’s well-being; removing this cornerstone would lead to society gradually falling apart. For characters in present-day Dutch literary fiction, this also seems to hold: removing the family roles between the 2,137 characters in the 170 novels would lead to the decline of almost 40% of the relational roles in the corpus, and arguably to less connected, more fragmented fictional populations.

The overrepresentation of the colleague role as opposed to the friend and lover roles might be indicative of the dominance of work-related social roles. This ties in to another commonplace: the separation of professional and personal life. These findings suggest that the societies depicted in the corpus tend to be more focused on the former than the latter.

While some broad patterns regarding the hierarchies of depicted social groups in present-day Dutch literary fiction can be deduced from the descriptive statistics reported in this chapter, a network analytic approach is required to specify the nature of the relations between these groups. Whereas a first insight into the relational dimension is already hinted upon by the distribution of relational roles, it is still unclear how and to what extent characters with certain demographic backgrounds interact with one another on the level of the text. The subsequent chapter on centrality presents an approach to extracting social networks of characters from each of the texts in the corpus and proposes and evaluates a model to compute the centrality of characters based on the structure of these fictional social networks.