A Distant Reading of Gender Bias in Dutch Literary Prizes

Noa Visser Solissa[1] and Andreas van Cranenburgh[1]

[1]University of Groningen

The Dutch literary scene has been criticized by authors for a lack of diversity and gender inequality. The two most important prizes, the *Boekenbon Literatuurprijs* and the *Libris Literatuur Prijs*, show this gender inequality, as about 80% of the nominated books were written by men, despite an equal author gender distribution among published literary books in the Netherlands. Given the over-representation of men in Dutch literary nominations, this inequality may be reflected in the word use of the authors, as people tend to use similar language as their peers. Therefore, this paper investigates whether it is possible to identify author gender inequality in Dutch literary prizes using distant reading techniques: text classification, topic modeling, and stylometry.

We collect a corpus of 300 literary books, divided into three categories: nominated (Nom), not nominated books written by a nominated author (NomAut), and books written by an author who has never been nominated (NotNom). A classification model trained to predict the category of a book reaches a cross-validated accuracy of 58.7%, surpassing the majority baseline (34%). Thus, nominated and not nominated books have distinctive textual features, which supports the view that literary quality is associated with particular formal features such as word usage. However, this word usage seems to be further removed from women writers, as the classification of books written by women consistently shows the lowest performance. The analysis of topics in the corpus suggest that the relation between nominated and not nominated books and author gender highly depends on the topic which is investigated. The difference in writing style of nominated and not nominated books cannot be clearly defined, but the results do suggest that the writing style of Harry Mulisch and Herman Koch may have influenced the writing styles of books nominated for literary prizes.

1. Introduction

Dutch authors have been criticizing the homogeneity and the dominance of white men in Dutch literary prize nominations and the Dutch literary scene (Amatmoekrim, 2015; Ramdas, 1997; Rouw, 2015; Weijers, 2014). This homogeneity is clearly seen in the Dutch literary prizes. In general fiction, men win substantially more literary prizes

than women. For the two most important prizes, the *Boekenbon Literatuurprijs* and the *Libris Literatuur Prijs*, 80% of the nominated books from 1987 to 2020 are written by men. Considering that an equal number of women and men publish novels in the Netherlands (Koolen, 2018), such a discrepancy is quite remarkable. Not only is the percentage of nominated books by men much larger than the percentage of books by women, but the percentage of men with multiple nominated books is also higher than for women. In 2020 the writers collective *Fixdit* (i.e., fix this) was founded to address the gender inequality and lack of diversity in the Dutch literary canon and scene (Fixdit, 2023). We investigate this inequality empirically with distant reading methods.

The *Libris Literatuur Prijs* acknowledged and analyzed the gender inequality in their nominations (Dijkgraaf and Appel, 2013). The results indicate that fewer women are nominated for the long list than expected from the number of books by women on the gross list. Over the last decade, more women have made up the majority of the jury members in the Dutch literary award scene (Boudewijn, 2020). However, juries with more women do not nominate more women writers (Dijkgraaf and Appel, 2013).

The dominance of white men in the Dutch literary scene is enforced by several factors besides literary prizes. Literary publishers and other professionals value formal aspects of literary works, and perceive prestigious novels as 'literary' and 'universal' (Koren and Delhaye, 2019). They often place white writers in the framework of 'literary' and 'universal' works. Contrarily, non-white writers and publishers are placed in frameworks based on their identity. For example, book reviews in Dutch news articles stress the ethnic and cultural background of non-white writers more, in comparison to German newspapers and newspapers from the USA (Berkers, 2009). This emphasis creates the idea that books written by non-white authors are different from the Dutch norm of literary quality, positioning these works outside of the norm (Staszak, 2009). Another factor that is likely to influence the inequality in the nominations of books is the influence of prestige of the genre, the author, and the books (Koolen et al., 2020; van der Deijl et al., 2019). Lastly, the homogeneous idea of literary quality is reinforced by the Dutch school curriculum. Dera (2021) shows that the majority of texts students read was written by Dutch white men. Women and non-western authors are structurally underrepresented in the curriculum, which upholds the idea that the norm of literary quality is associated with white, western men (Dera, 2020).

The association between literary quality and white, Western men is upheld by multiple factors, such as the identities emphasized in book reviews and the manner in which school curricula teach students what literary quality is. As literary awards are supposed to award the 'best' book, it is interesting to further investigate how the texts themselves relate to this homogeneous norm of literary quality, by researching the word use and topics within nominated and not nominated books.

Homogeneity in the literary scene does not only exist in the Netherlands. Several projects have been set up worldwide to quantify the gender breakdown of major literary works and book reviews, such as Stella,[1] focused on Australian writers, the VIDA count,[2] focused on the United States of America, and Frauen Zahlen[3] and Literaturkritik in Zahlen,[4] both focused on books written in German. These projects focus on publications and book reviews, as book reviews, particularly in major newspapers,

1 https://stella.org.au/initiatives/research/.
2 http://www.vidaweb.org/the-count/.
3 http://www.frauenzÃđhlen.de/.
4 https://www.uibk.ac.at/iza/literaturkritik-in-zahlen/.

have a large influence on the popularity of a novel. Unfortunately, the outcomes from these projects show that books written by women do not receive the same attention in major newspapers and book reviews as books written by men. Thus, (white) men seem to dominate the literary scene of several Western countries, as these are the books that are most often read, reviewed, and nominated.

Although this paper will focus on author gender inequality, it is important to note that all the authors mentioned concerning the *Libris Literatuur Prijs* and *Boekenbon Literatuurprijs* are white, as other forms of inequality, such as ethnic and cultural background, also lead to a form of homogeneity in the Dutch literary scene. Due to limitations of the corpus available, other forms of inequality besides author gender could not be investigated. Additionally, the analysis of author gender will only focus on men and women, again due to the limitations of the dataset.

It is clear that the causes of the homogeneity in Dutch literary awards are multi-faceted, which will be further discussed in Section 2.3. Given that there is an over-representation of white men in Dutch literary nominations, this inequality may be visible in the word use of the authors, as people tend to use similar language as their peers (Eckert, 2012).

Therefore, this paper investigates whether it is possible to identify author gender inequality in Dutch literary prizes using distant reading methods. We will do so by answering the following three research questions:

RQ1 To what extent can nominated and not nominated books be distinguished based on textual features alone?

RQ2 Is there a relation between classifications of nominated versus not nominated books and author gender?

RQ3 Are the differences in topics and writing styles between books that are nominated for literary prizes and those that are not related to author gender?

The goal of the first question is to investigate whether it is possible to identify nominated and not nominated books based on textual features using a classification task, which has not been researched before. The second question aims to explore if a relation between nominated books and author gender can be identified using textual features. The goal is to relate the results of the author gender classification to nominated and not nominated books, and to analyze how these patterns relate to the results of the model trained to classify nominated books. The last question aims to identify the topics in nominated and not nominated books using unsupervised algorithms. The goal is to explore which topics occur more in nominated books, and are therefore probably associated with higher literary quality. For the writing styles, specific word use related to nominated books and not nominated books will be identified. These results will be used to give a more interpretable insight into the relation between nominated and not nominated books and author gender. This paper builds on the research on nominated novels of Koolen and van Cranenburgh (2017) and books from that research will therefore be used in our corpus.

We hypothesize that nominated and not nominated books can be identified based on word use. We also hypothesize that, due to the dominance of men in literary nominations, nominated books written by men will be easier to classify compared to nominated books by women; and vice versa for not nominated books.

2. Theoretical Framework

2.1. Computational stylistics & literature

Computational literary studies, and computational stylistics in particular, is a field that focuses on modeling 'literary discourse' using computational and statistical methods (Herrmann et al., 2021). It can be grouped into three categories: formalist, social and cognitive approaches (Herrmann et al., 2021). Formalist approaches focus on understanding the distinctive features and structures of literary works, including the manner of writing that constitutes literariness, the nature of genres, literary quality or authorial style. Social approaches investigate social practices across communities, such as 'canonicity' and 'prestige.' Cognitive approaches research the 'cognitive' side of aesthetics and stylistics, such as the psychology of literature and reader response.

Formalist approaches Writing style can be seen as a complex system of combinations of formal features (Herrmann et al., 2021), in which formal features are linguistic features on character, lexicon, syntax and semantic level. The most reliable features for measuring stylistic similarity and distinction are function words (Burrows, 2002). Computational stylistics is based on the assumption that individuals have idiosyncratic and largely unconscious habits of language use, leading to stylistic similarities between texts written by the same person (Evert et al., 2017). Therefore, computational techniques can determine authorship, due to the relative frequency of function words, parts of speech, degrees of vocabulary richness or syntactic complexity (Lupei et al., 2020; Marsden et al., 2013; Tuzzi and Cortelazzo, 2018; Varela et al., 2016). Different authors use measurably distinct styles by over-utilizing or avoid particular common words and phrasing, despite using the same structural and grammatical bounds of a common language (Marsden et al., 2013). Writers favor (or filter) certain words in a manner which goes beyond the use (and avoidance) of common phrases due to word use in social groups. This word preference creates an individual style which can be identified probabilistically. Thus, computational techniques lend themselves for identifying distinctive writing styles.

Social approaches One of the key areas of the socially-oriented frameworks in computational stylistics is examining the relationship between representation and inequality, by examining inequalities and biases of representation in literary and other cultural documents (Herrmann et al., 2021). Representations are explored on two levels, namely on the level of agents, such as authors and characters, and on the level of form, such as style and semantics.

An example of representation on agent level is Underwood et al. (2018), which shows a massive decline of women authors in English fiction in the twentieth century. The form level of representation is also explored in this study, as the historical investigation of English fiction in the twentieth century also showed that the gender division between characters becomes less sharply marked over this period of time, suggesting a growing equality in gender representation in characters.

Lejun et al. (2021) is one of the few studies on the relation between the representation of characters and literary prizes on the level of form. They found that a high concentration of characters and emotion fluctuation are common characteristics in works by authors nominated for the Nobel Prize in Literature in 2012 and 2013. As the concentration in which characters are mentioned and the manner in which emotions are expressed are ways of portraying characters in novels, these results suggest that the

manner in which characters are portrayed in novels can be related to the nomination for literary prizes and the perception of literary quality.

On the agent level, van der Deijl et al. (2016) have shown that Dutch authors write predominantly about characters close to their daily life experience. As the authors are predominantly men, this results in an over-representation of men as main characters in Dutch literature. They also show that the narrating main characters are predominantly highly educated men of Western descent, similar to the majority of the authors in the corpus. Authors also appear to portray characters of different genders in very different professional settings. In the corpus, student is the most common occupation for both male and female characters. However, for men, the third and fourth most common professions are entrepreneur and teacher, whereas for women those are sex worker and housewife.

Thus, the homogeneity of author gender in Dutch literature seems to influence the way men and women are described in novels. In addition, the homogeneity in the characters of literary novels, and the manner in which the characters are portrayed, could have an effect on whether a novel is perceived as literary or not. Smeets et al. (2019) provides a more nuanced take on these results, as their social network analysis on Dutch characters in contemporary novels shows that women and immigrant characters statistically take up a more central position in these novels than men and non-immigrant characters. Due to the limitations of their corpus, these results could be skewed. Therefore, they argue that future research should strongly connect qualitative and quantitative strands.

2.2. Textual features & literary quality in Dutch literature

As this paper investigates Dutch literary books, a more in-depth overview of computational stylistics research on literary quality in Dutch literature will be given. In order to investigate author gender inequality in Dutch literary prizes, it is important to understand how literary judgments on Dutch literature can be predicted using computational techniques. To do so, van Cranenburgh and Bod (2017) used the results of the National Reader Survey, which measures the perception of literary quality by Dutch readers. In this survey, readers could rate books on literary quality, both on books that the respondents had read and books that they had not read (Koolen et al., 2020). For the books that the readers had not read, respondents could fill in the rating of literary quality they expected the book to have. The two main motivations given by the respondents to rate literary quality were genre and the text itself.

The results show that respondents base their expectations of literary quality on literary quality from the 'genre' of the book, such as suspense and chick lit (van Cranenburgh and Bod, 2017). Detectives, thrillers and chick lit are not perceived to be of high literary quality, whereas literary novels are mainly perceived to be of high literary quality. This influences the rating of women writers, as these books are more often marketed within a particular, gendered, genre. This relation between author gender and genre is in line with the findings of van der Deijl et al. (2019), which show a clear relation between certain genres and author gender in online literary communities. From the difference in literary ratings between novels of different genres, as well as the motivations given by the respondents, Koolen et al. (2020) conclude that a consensus of literary quality exists among Dutch readers. This consensus is grounded in textual features such as writing style, which includes sentence length and word usage.

The results of The National Reader Survey have also been analyzed computationally with respect to the text of the novels in the corpus. Van Cranenburgh and Bod (2017)

use the results of The National Reader Survey to predict the average literary rating in the survey based on textual features. The results show that the literary quality of Dutch novels can be predicted from textual features alone to a substantial extent. Secondly, the results show that it is important to use novels in original language only when analyzing author gender.

A drawback of the National Reader Survey and the research based on its results, is that author gender is not evenly distributed across genres in the corpus. Despite the fact that this corpus of 401 Dutch novels has an almost equal percentage of men and women writers, this is not seen in the subset of general fiction. In this genre, there are more originally Dutch works by men, and more translated works by women. As genre and author gender both influence literary ratings, Koolen and van Cranenburgh (2017) analyzed a second corpus of general fiction, consisting of an equal amount of works from women and men. They show that it is possible to use topic modeling to investigate and interpret how topics in novels relate to author gender. For example, the topic 'military' is strongly related to works by men, whereas the topic 'settling down' is strongly related to novels by women.

Thus, previous research suggests that it is possible to distinguish nominated and not nominated books based on textual features, as it seems possible to predict literary quality based on textual features (van Cranenburgh and Bod, 2017; van Cranenburgh and Koolen, 2020). It is important to take genre and author gender in account in this type of research, and to use interpretable models to draw nuanced conclusions and limit the reproduction of stereotypes (Koolen and van Cranenburgh, 2017).

2.3. Gender as a social variable

In Natural Language Processing (NLP), gender is often treated as a biological characteristic. This is a very limiting view of gender, as it ignores the agency of a speaker. This view also goes against gender theory and social science, where it is considered that gender is something that someone *does* instead of *is* (Nguyen et al., 2016). Additionally, individual language use varies due to the social group someone is situated in or communicates with (Eckert, 2012). As peer groups are often homogeneous in gender and age, people of the same gender and age have a language use that is more closely related to each other. Thus, the relation between gender and language is social.

As this article focuses on the relationship between author gender and nominations for literary prizes, it is important to clearly define how the variable gender will be used throughout this paper. Gender is an ethically complex feature to use in NLP research, as it is a social construct (Butler, 1998). It is often implemented as a binary variable, whereas more than two gender identities exist. Keyes et al. (2021) argue that it is important to treat gender as 'multiplicitous': a concept which has many meanings and relations to individuals and communities.

Recent NLP research has also argued that gender should be approached as a social variable, rather than a static biological one (Bamman et al., 2014a; Nguyen et al., 2014). As language is inherently social, individual speakers often diverge from the gender stereotypes that are found in many studies (Nguyen et al., 2016). Even though certain language features are used more by a certain gender on average, NLP research should refrain from drawing generalizing conclusions. Furthermore, gender varies in different cultures and languages, and linguistic variation can also be identified among speakers of the same gender.

Argamon et al. (2003) investigated the difference between the writing of men and women, in English fiction and non-fiction. They show that differences in writing style

are seen between authors or different genders, and that these differences are strongly related to genre. They find that the writing style of women aligns more with fiction, whereas the writing style of men is more related to non-fiction. Based on the distinctive features found, they conclude that women write in a way that is more 'involved,' while men write in a manner that is more 'informative' (Biber and Finegan, 1989). Argamon et al. (2003) argue that the gendered difference between 'involved' and 'informative' writing is due to the differences in socialization of people of different genders. They also argue that the significant relation between gender on the one hand, and fiction and non-fiction on the other, is related to the cultural situation that the genres are placed in. However, Argamon et al. (2003) do not specify in what way the cultural situation of the texts they explored is gendered.

To isolate the influence of genre on (gendered) writing style, Herring and Paolillo (2006) investigate the influence of gender in language, when the genre of text is constant. They analyzed weblogs of two different genres: diary and filter. The 'diary' blogs report on the author's life, while 'filter' blogs report on events external to the author's life. Surprisingly, no significant correlation between the stylistic features and gender was found. Significant correlations were found between woman preferential features and personal blogs and man preferential features and filter blogs. Thus, they conclude that genre is a stronger predictor than author gender of the 'gendered' stylistic features found by Argamon et al. (2003). They argue that genres appear to be gendered, due to the topics discussed. They also hypothesize that men and women use similar language within a genre, and that therefore influence of gender on language is not identified within one genre. Thus, it is important to carefully draw conclusions on gendered language use, as gendered language use can be strongly related to the topics within a text (Herring and Paolillo, 2006; Koolen and van Cranenburgh, 2017).

Influence social group on (gendered) word use Bamman et al. (2014a) used clustering to analyze how differences in word use between genders relate to the topics in Tweets and to the social network of individual Twitter users. They predicted the author gender of 14.000 Twitter users and reach an overall accuracy of 88% in binary gender prediction; author gender can therefore be accurately predicted using only word features. In addition to a binary author gender classification, they clustered the Twitter users based on their tweets to find a more natural grouping of writing styles and topics. The clusters show multiple expressions of gender, such as interactions between gender and age or race, underlining the importance of intersectionality. The clusters are also related to certain topics, such as athletes and sport-related organizations. From these topics, Bamman et al. (2014a) conclude that in their data, men are more likely to write about hobbies and careers. As these topics are related to large numbers of named entities, men use more named entities in their language. They state that these specific topics are the most probable explanation for the usage of named entities by men, and not 'informativity' or 'explicitness'. Lastly, Bamman et al. (2014a) analyzed the relation between author gender and writing style using the social network of Twitter users. They found that users who have a social network that includes fewer same-gender social connections, use language that is not matched with the classifier's model for their gender, and vice versa. For example, the segment of women which have been classified as women with strong confidence, have an average network composition that consists of 77% women.

Another approach in which the multifaceted positioning of language use is shown is in interactive research. Nguyen et al. (2014) implemented an online game, in which

players guessed the gender and age of a Twitter user. The results suggest that 10.5% to 16% of the Dutch Twitter users do not use language corresponding with language the players expected to be used by people of the users' gender. To analyze this further, a gender continuum was created, using the percentage of players who guessed the user to be a man and the percentage of guesses for woman were calculated per Twitter user. This showed that the guesses of the players were based on the expected linguistic behaviour of women and men. It also showed that the distribution of percentages of players that guess man and woman cannot be grouped into two distinct groups. These results underline not only that gender should be treated as a social variable, but also that the influence of gender on language use and perception is limited and nuanced.

From the results of Bamman et al. (2014a) and Nguyen et al. (2014) it does not seem that women communicate in a manner that is more 'involved' or that men communicate in a manner that is more 'informative' due to socialization (Argamon et al., 2003), but rather that people communicate and expect other people to communicate in a certain way, based on the social group that they are communicating with, such as Tweets focused on a (gendered) topic (Bamman et al., 2014a). As Nguyen et al. (2014) did not find that distinctive gendered groups guessed a certain gender per Twitter user, it seems that the perception of gendered language use is not related to the gender of the guesser either.

3. Method

We collect a corpus of 300 original Dutch literary books from 1989–2012 (see Appendix A for the full list of books).

The corpus is divided into three subcorpora:

1. Nom: nominated books,
2. NomAut: not nominated books by nominated authors, and
3. NotNom: not nominated books by not nominated authors.

The NotNom books are published by the same publishers as the Nom books, and were selected to resemble the same distribution of publication years as the Nom books. The distribution between these three categories, author gender distribution and number of unique authors can be found in Table 1. The average number of words per sentence and average total length of the Nom, NomAut and NotNom books can be found in Table 2.

The majority of the Nom books have been nominated for the *Libris Literatuur Prijs*, since the long lists for this prize are publicly available for the years 2005-2020. For the nominations of the *Libris Literatuur Prijs* before 2005 and the *Boekenbon Literatuurprijs*, the long lists have not been made public. Sixteen books in the Nom subcorpus have been nominated for both the *Boekenbon Literatuurprijs* and the *Libris Literatuur Prijs*.

It should be noted that not all the books in the corpus are novels, since not all nominated books are fictional, such as *Congo* by David van Reybrouck. Therefore the corpus contains novels, essays, histories, novellas, op-eds, poetry, and a (rather free) Quran translation.

To answer the three sub-questions, three different NLP techniques are used, namely text classification, topic modeling, and stylometry. The first technique, text classification, is a supervised approach, and the latter two are unsupervised. We use the text classification experiments to derive quantifiable conclusions, whereas topic modeling

	Nom	NomAut	NotNom	Total
Books	100	102	98	300
Unique authors	73	35	83	191
Books by women	36	42	43	121
Books by men	64	60	55	179

Table 1: Number of books in the three subcorpora

	Whole corpus	written by women	written by men
Words per sentence			
Nom	13.18	13.32	13.10
NomAut	12.61	11.97	13.06
NotNom	13.51	13.25	13.72
Words per book			
Nom	62712.9	61787.8	63233.3
NomAut	71559.0	72677.6	70776.0
NotNom	78996.5	79217.8	78823.5

Table 2: Mean words per sentence and book.

and stylometry were used for more qualitative interpretations of the results obtained with text classification.

The text classification results are used to answer RQ1 and RQ2, and the results of topic modeling and stylometry for RQ3. Due to the different objectives for the techniques used, the research designs will be discussed per technique.

3.1. Text Classification

A well-established method to analyze text in relation to a variable of interest such as literary prestige or author gender is text classification with logistic regression using word frequencies (Bamman et al., 2014a; Fast et al., 2016; Herring and Paolillo, 2006; Koolen and van Cranenburgh, 2017; Nguyen, 2017). In this research, we consider three classification tasks:

1. identifying Nom, NomAut and NotNom books; see Table 3,
2. identifying whether a books has been nominated (Nom) or not (NomAut and NotNom), see Table 4; and lastly
3. identifying author gender; see Table 5.

We use the counts of the 5000 most frequent words and bigrams as features, and normalize the counts with Tf-Idf to give more weight to distinctive features. Previous work already indicated the effectiveness of such bag-of-word features for predicting literary judgments (van Cranenburgh and Bod, 2017) and author gender (Bamman et al., 2014b). We use regularized logistic regression and evaluate using 5-fold cross-validation. The cross-validation folds are stratified by author, such that the model is never trained and tested on books from the same author, which would enable the model to pick up on author style as a shortcut to a correct classification. The predictions are evaluated using precision, recall, F1-score and overall accuracy. Precision is the

fraction of correct predictions in a particular class of the total predictions made of that class. Recall is the fraction of correct predictions of a particular class of all instances in the target class. F1-score is the harmonic mean of precision and recall.

In addition, it is possible to inspect the confidence of each classification made by the logistic regression model. This confidence is expressed as a probability that the classifier assigns to a given label for a text.

3.2. Topic Modeling

A topic model is an unsupervised technique for summarizing a corpus using automatically identified topics. Latent Dirichlet Allocation (LDA) is an iterative probabilistic topic model (Blei et al., 2003). LDA automatically assigns topics to documents, and words to topics, in the form of probabilities. The probabilities are updated iteratively, with the objective of summarising each document with a small number of relevant topics, and each topic with a small number of relevant words.

LDA is used in a wide range of different fields, such as Twitter-analysis, biomedical science and literature (Jelodar et al., 2019). One of the advantages of using LDA to analyse literature is that it can reveal patterns that are not easily observed. For example, when analysing changes in literature over time, LDA can identify patterns that are not easily recognized because they happened gradually, or simply have slipped under the radar (Goldstone and Underwood, 2014). The same authors also argue that another advantage is that the unsupervised topics and clusters created force researchers to analyze literature outside of predefined, traditional concepts. A disadvantage of LDA topic modeling is that researchers overestimate their ability to explore large corpora quickly, despite the fact that the topics created might not be as coherent and stable as they seem (Schmidt, 2012). The set of words related to one topic do not by definition have anything in common except for common co-occurrence. Thus, if a topic occurs in two different documents, it does not necessarily mean that this particular set of words related to a topic has the same relation to that topic within these two documents. Therefore, the interpretation of LDA topic modeling is sensitive to the interpretation of the researchers, and conclusions should be carefully drawn from it.

3.3. Stylometry

To investigate the difference in writing style between books that have been nominated and books that have not been nominated more closely (RQ3), we use Cosine Delta to identify the difference in writing styles between literary books that have been correctly classified in the nominated or not and author gender classifications, and books that have been misclassified in all these classifications.

Cosine Delta is a successful technique to identify authorship and writing style using the most frequent words of books (Evert et al., 2017). The model shows which books in the corpus have a similar writing style, and how the different writing styles of the books relate to each other. Therefore, a comparison between the consistently correctly classified and misclassified books is chosen, as the books that have been correctly classified in all models have a word use that is consistently related to the features related to nominated books (Nom), not nominated books by nominated authors (NomAut) and not nominated books by not nominated authors (NotNom) classes and author gender. The misclassified books have a word use that is clearly hard to relate to the features related to their target classes. Thus, it can be expected that the most clear distinction in writing style can be found between these sets of books. This comparison

will predominantly be used to obtain an indication on the distinctive writing style that is related to nominated books and to attempt to relate this distinctive writing style to author gender.

Whole corpus	Precision	Recall	F1-score	# books
Nom	56.2	69.0	61.3	100
NomAut	56.2	35.3	43.4	102
NotNom	64.0	72.4	67.9	98
Overall			58.7	300

Women	Precision	Recall	F1-score	# books
Nom	46.5	55.6	50.6	36
NomAut	53.8	33.3	41.2	42
NotNom	65.4	79.1	71.6	43
Overall			56.2	121

Men	Precision	Recall	F1-score	# books
Nom	59.8	76.6	67.1	64
NomAut	57.9	36.7	44.9	60
NotNom	62.7	67.3	64.9	55
Overall			60.3	179

Table 3: Logistic regression results on nominated books (Nom), not nominated books by nominated authors (NomAut) and not nominated authors (NotNom). The second and third table show a breakdown of the results per author gender, man or woman. The majority baseline for this binary classification task is an accuracy of 34%.

Whole corpus	Precision	Recall	F1-score	# books
Nominated Books	59.6	62.0	60.8	100
Not Nominated Books	80.6	79.0	79.8	200
Overall			73.3	300

Women	Precision	Recall	F1-score	# books
Nominated Books	54.8	47.2	50.7	36
Not Nominated Books	78.9	83.5	81.1	85
Overall			72.7	121

Men	Precision	Recall	F1-score	# books
Nominated Books	61.6	70.3	65.7	64
Not Nominated Books	82.1	75.7	78.7	115
Overall			73.7	179

Table 4: Logistic regression results: nominated (Nom) or not nominated (NomAut and NotNom). The second and third table show a breakdown of the results per author gender, man or woman. The majority baseline for this binary classification task is an accuracy of 66.6%.

Whole corpus	Precision	Recall	F1-score	# books
Man	73.1	80.4	76.6	179
Woman	66.0	56.2	60.7	121
Overall			70.7	300

Nom	Precision	Recall	F1-score	# books
Man	75.7	87.5	81.2	64
Woman	69.2	50.0	58.1	36
Overall			74.0	100

NomAut	Precision	Recall	F1-score	# books
Man	69.6	80.0	74.4	60
Woman	63.6	50.0	56.0	42
Overall			67.6	102

NotNom	Precision	Recall	F1-score	# books
Man	74.1	72.7	73.4	55
Woman	65.9	67.4	66.7	43
Overall			70.4	98

Table 5: Logistic regression results: author gender. The results on the whole corpus are also broken down by nomination class, Nom, NomAut or NotNom. The majority baseline for this binary classification task is an accuracy of 60.0%.

4. Results

4.1. Text Classification

The text classification results show to what degree nominated and not nominated books can be distinguished based on textual features alone. The two models trained to identify nominated books (Nom, NomAut versus NotNom, and nominated versus not nominated) both obtain an accuracy higher than chance or the majority baseline. The results also show that the not nominated books by nominated authors (NomAut) are the hardest to classify, since these scores are the lowest. The results in Table 3 and 4 also show that for the books by women, not nominated books have the highest scores. For books by men, the Nom books have the highest score in the Nom, NomAut, NotNom model. Also, the difference in F1-score between the nominated and not nominated books is larger for the books written by women than for the books written by men.

The logistic regression predicting author gender confirms this pattern. The books written by women score consistently lower than the books written by men. The difference in F1-score between these author genders are, however, smallest for the NotNom books. Thus, the results seem to indicate a relation between the word use in books written by women and not nominated books by not nominated authors, as the books written by women consistently have the highest score for the NotNom class. For the books written by men, such a relation between the Nom, NomAut and NotNom classes was not found, but the books written by men did consistently have higher results than the books written by women, for all classes. This was probably not due to the higher number of books by men in the corpus, as this pattern was also seen, though not as strongly, in models trained on a gender-balanced subset of the corpus.

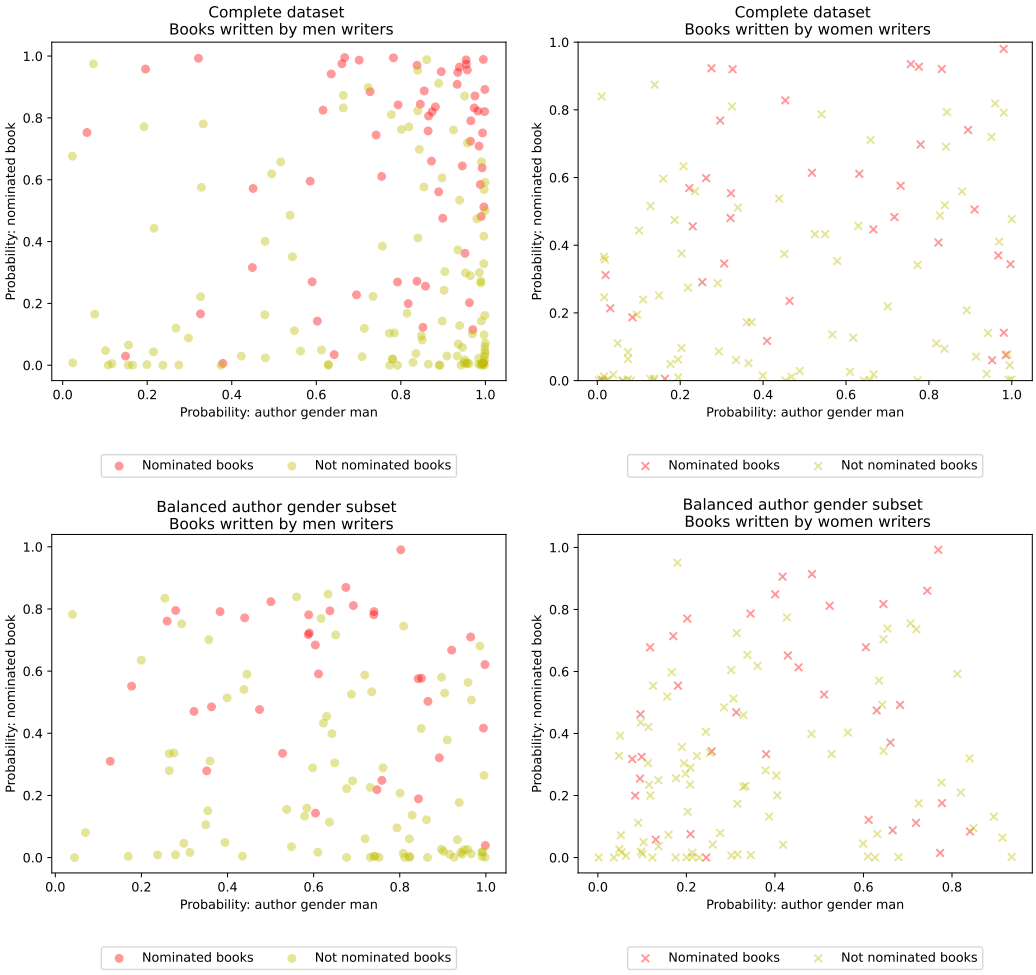

Figure 1: Probability of nominated-or-not and author gender logistic regression models that a book is written by a man, and that a book has been nominated. A probability higher than 0.5 means that the classifier predicts that a book was written by a man or was nominated.

Confidence of classification Figure 1 shows that for the nominated-or-not model, there seems to be a relation between a high probability to be nominated for a literary prize, and books written by men. This is shown by the few books written by men that have a low probability of being written by men and a high probability of being written by a woman. For the books written by women, relatively more books have a high probability of being written by men and have won a literary prize. Also, the probability of being nominated for a literary prize in general is lower for books written by women than for books written by men.

The confidence of these two classification tasks is shown per author gender and per dataset (complete dataset and balanced author gender subset). As can be seen, the confidence for books written by men ranges from 0.0-1.0, and as expected, the majority of the nominated books (87.5%) have a confidence higher than 0.5 to be written by a man. This pattern is seen in the balanced author gender subset as well. For the books written by women, there is not such a clear skew towards the left side of the x-axis. Furthermore, it should be noted that 41.7% of the nominated books written by women are predicted to be written by a man. This is interesting, because it would mean that books written by women are more likely to be predicted as written by a man.

In the plots of the books written by men, few books are predicted to be written by a woman and nominated for a literary prize. Also, the nominated books all have a high probability of being written by men. Only 12.5% of the nominated books written by men are predicted to be written by a woman. Thus, it seems that nominated books by men have a high probability of being written by men according to the nominated-or-not model.

For the plots of the books written by women, it is noticeable that only two books have a probability of being nominated that approach 1 in the complete dataset. Also, only 19.0% of the books have a probability higher than 0.6 to be nominated, which shows that only this percentage of the books written by women are predicted to have been nominated with high confidence. These results show that in general, books written by women overall have a lower probability of being nominated according to the nominated-or-not model. Similar results are seen in the balanced author gender subset.

To conclude, Figure 1 shows that there are few books which have a high confidence to be nominated and a high confidence to be written by a woman. This is in line with the results discussed in Table 3 and 4. Additionally, the difference in F1-scores for Nom written by women is lowest when classifying on author gender (see Table 5), suggesting that nominated books written by women are the hardest to distinguish from books written by men.

4.2. Topic Model

In order to interpret these results in relation to writing styles and topics of the books, we made an LDA topic model of the corpus creating 50 topics. The topic model resulted in 29 interpretable topics. An example of an interpretable topic is art (topic 3), including words such as: *schilderij* (painting), *schilder* (painter) and *kunst* (art). An example of an uninterpretable topic is topic 2, which consists of the words: *jaar* (year), *vriend* (friend), *student* (student) and *foto* (photo). Some topics, such as topic 29, are clearly related to certain books in the corpus. Topic 29 is about *Congo* by David van Reybrouck. Previous results with topic modeling of Dutch literature have found similar topics, including author- and book-specific topics (Jautze et al., 2016).

The topics have been analyzed by type of book, and can be seen in Figure 2 and 3. In

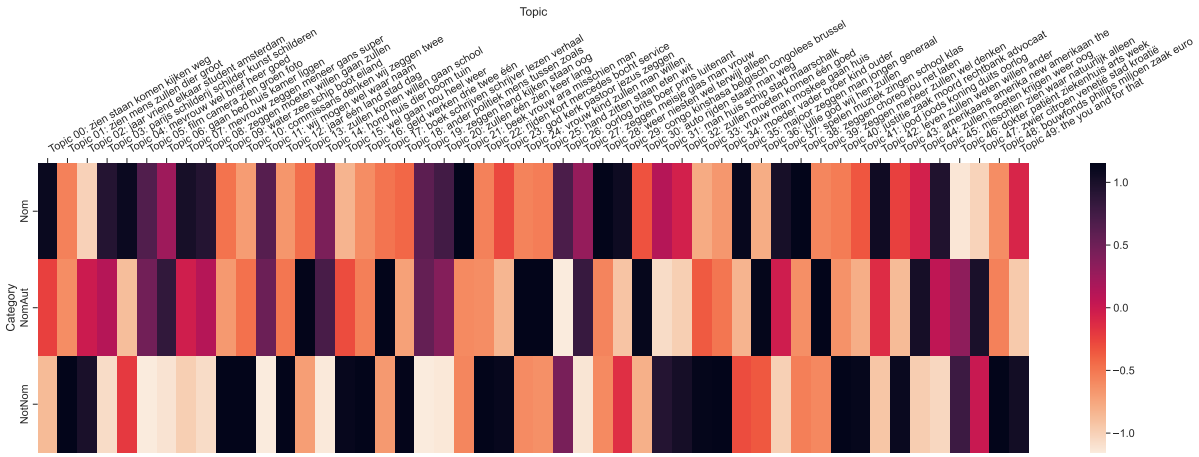

Figure 2: Heatmap showing the correlation of topics in Nom, NomAut and NotNom books. Dark colors indicate a strong correlation between a topic and that class, a light color indicates a weak correlation.

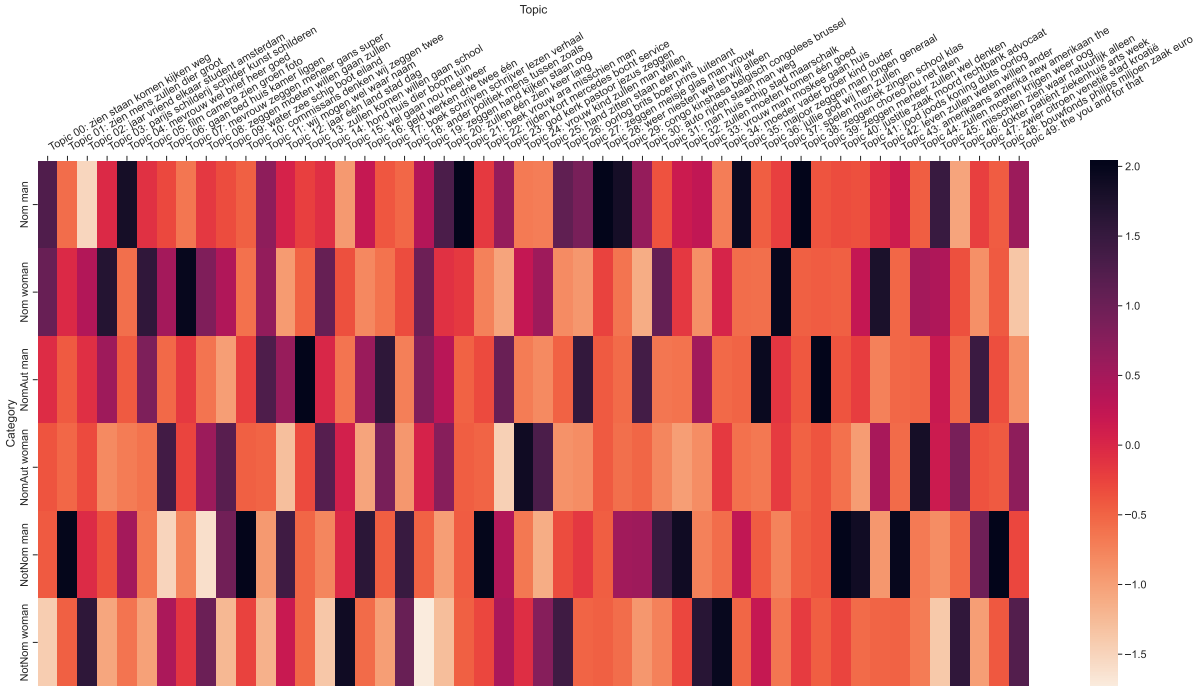

Figure 3: Heatmap showing the correlation of topics in Nom, NomAut and NotNom books, by author gender. Dark colors indicate a strong correlation between a topic and that class, a light color indicates a weak correlation.

Figure 2, the topics per Nom, NomAut and NotNom books are shown. The heatmap shows the average proportion of topics in these three classes, relative to each other. A few topics could be identified that are more typical in nominated or not nominated books, which could sometimes be related to author gender. For example, topic 41 'Second World War' in nominated books and topic 30 'driving a car' occurs most in not nominated books of nominated authors. Topics 31-34 are more strongly related to NotNom books, including a topics about the Islam and family.

When the topics are split by author gender, certain topics related to one of the three classes remain. For example, topic 0 'being on the road' strongly related to Nom books, for both men and women writers. Topic 17 'writing' occurs relatively more in NomAut books, by men and women writers. For the NotNom books, such topics cannot be defined.

Other topics seem to relate to a specific nomination class, but are actually more gender specific, such as 'driving a car' (topic 30). Based on Figure 2, it seems that this topic is mostly related to NomAut books. However, specifying the analysis on author gender shows that this topic is positively related to Nom, NomAut and NotNom books written by men, but most frequently in NomAut books written by men (see Figure 3). Another example is topic 6 'going home/sleeping' which is most strongly related to Nom, NomAut and NotNom books written by women. Lastly, there are also topics that are related to authors of a certain gender in particular classes. For example, topic 49 'English words' occur most in Nom books written by men and not nominated books (both NomAut and NotNom) written by women.

To conclude, Figure 2 show that there are certain topics that relate to Nom, NomAut and NotNom books specifically. The relation is complex, however, as some topics are related to a class due to author gender. For example, the topic 'driving' is related to NomAut books, but actually predominantly occurs in books written by men. Other topics identified by the topic model, such as the usage of English words, seem to be related to nominated books when a book is written by a man, and to not nominated books when a book is written by a woman.

4.3. Stylometry

We also used Cosine Delta to explore similarities and differences in writing styles between books that have been correctly classified in all three logistic regression classifications and books that have been misclassified in all three classifications. For this we consider the frequencies of the 3000 most frequent words (MFW) across the corpus.

Figure 4 and 5 show the cosine similarity of the writing style of the misclassified books and the correctly classified books. The rows representing the correctly classified books are either more blue or more red, which shows that the writing style in that book either is similar or dissimilar to the misclassified books. The heatmap for NotNom books is not included as only three NotNom books were consistently misclassified, which is not enough to draw a conclusion from.

For the Nom and NomAut plots (Figure 4 and 5), a pattern is seen where a specific author relates more to one author, and less to another. Thus, the writing style of each author relates differently to the misclassified books. For example, in the Nom plot, the misclassified book of Bianca Stigter is close in writing style to the correctly classified book of Joost Zwagerman. Additionally, there are a couple of nominated authors that are similar to all misclassified books. For example, for the correctly classified Nom books, these are books of Christiaan Weijts, Arnon Grunberg, A.F.Th. van der Heijden, and Herman Koch. For the NomAut books, examples of such books are works by Joost

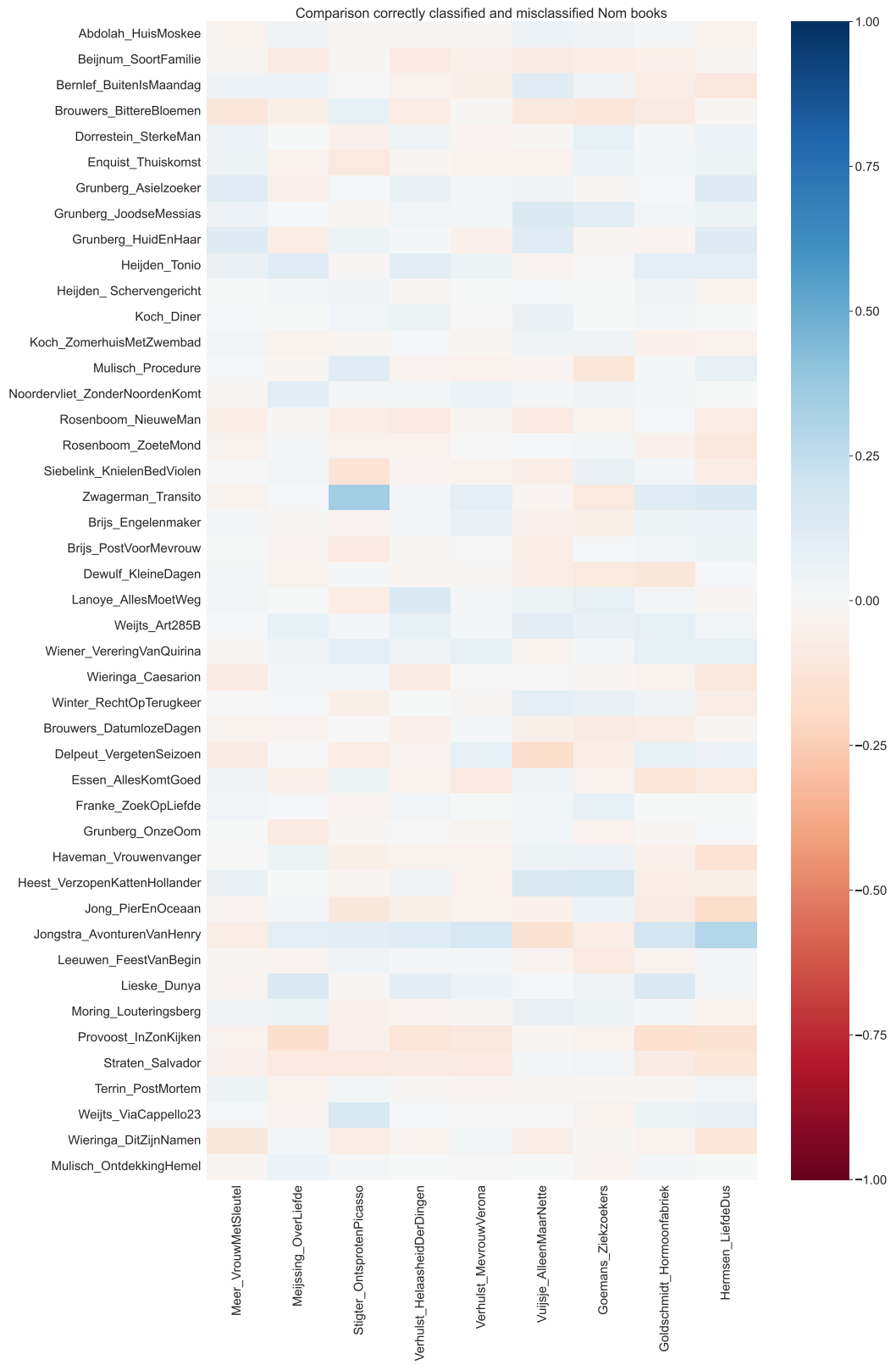

Figure 4: Heatmap showing the cosine similarity of the writing styles of the correctly classified books (rows) and the misclassified nominated books (Nom; columns). Blue indicates similarity between a topic and that class, while red indicates dissimilarity.

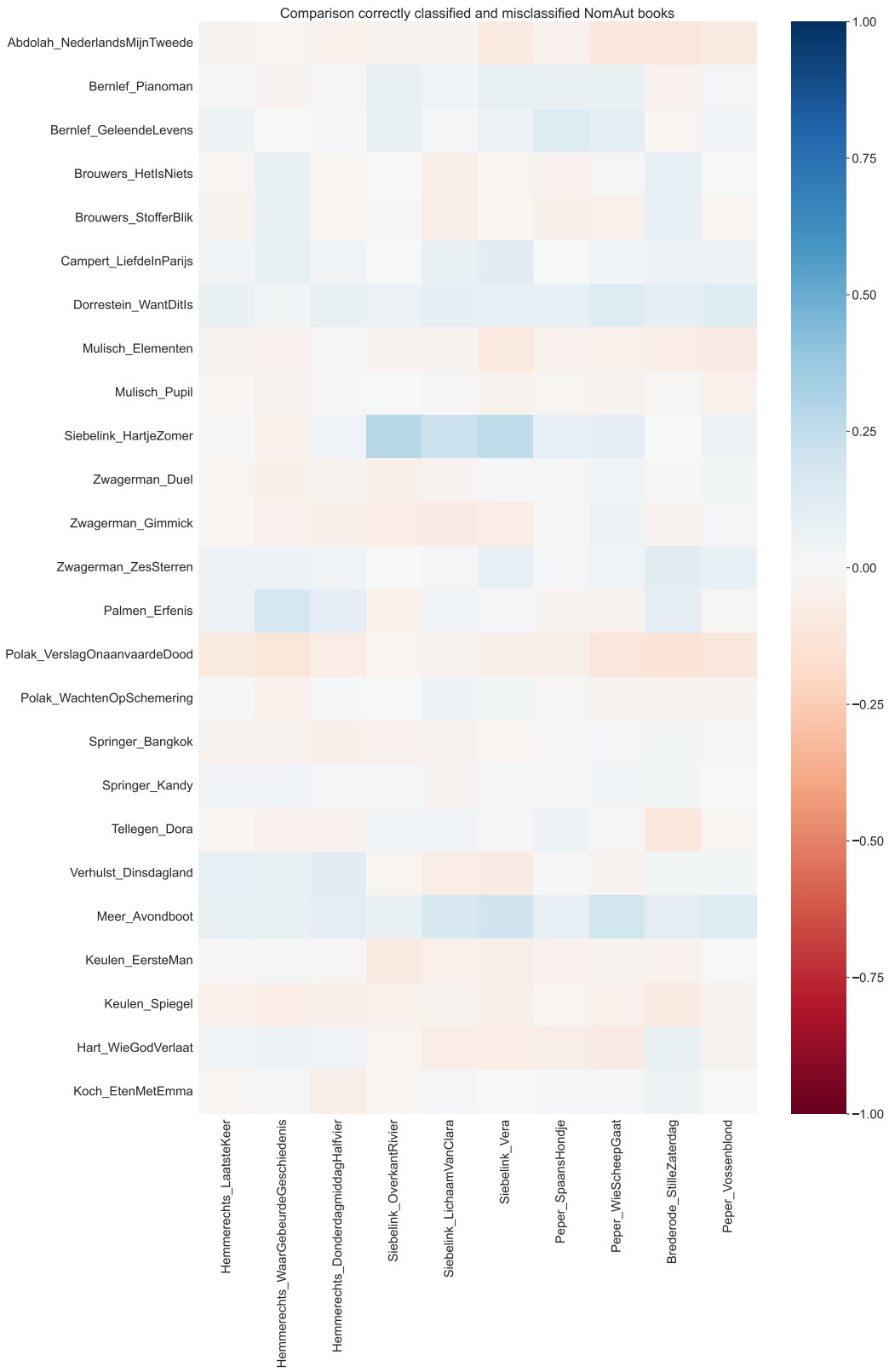

Figure 5: Heatmap showing the cosine similarity of the writing styles of the correctly classified books (rows) and the misclassified nominated books (NomAut; columns). Blue indicates similarity between a topic and that class, while red indicates dissimilarity.

Figure 6: PCA showing the relative positions of the correctly classified nominated books writing styles (Nom). The books written by men are blue and the books written by women orange.

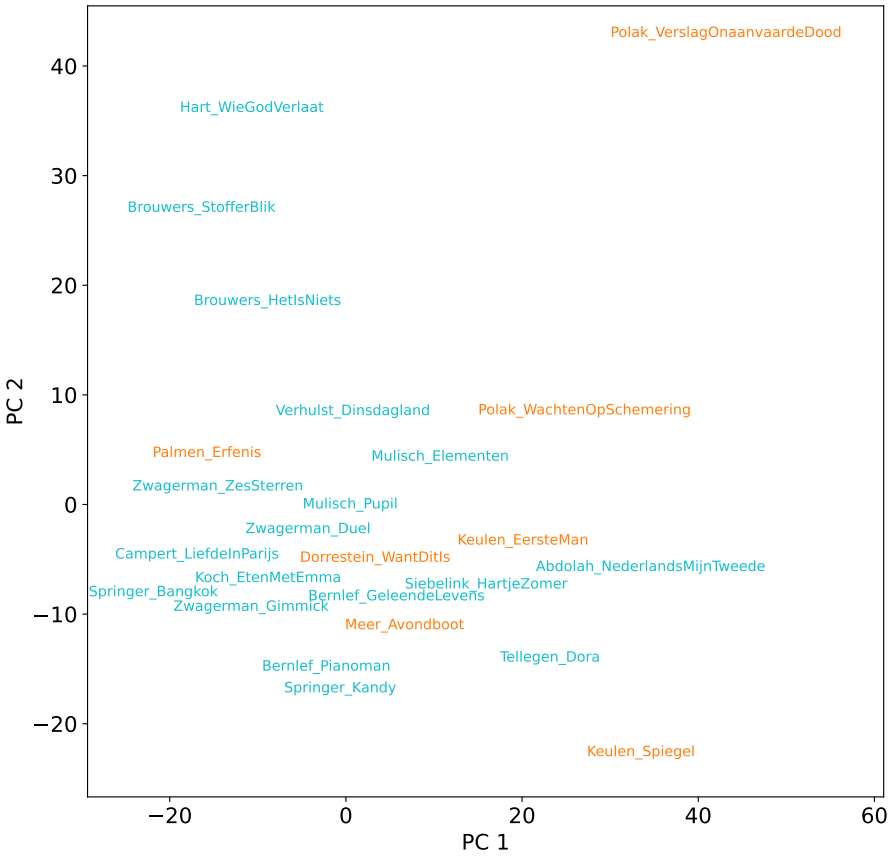

Figure 7: PCA showing the relative positions of the correctly classified nominated books writing styles (NomAut). The books written by men are blue and the books written by women orange.

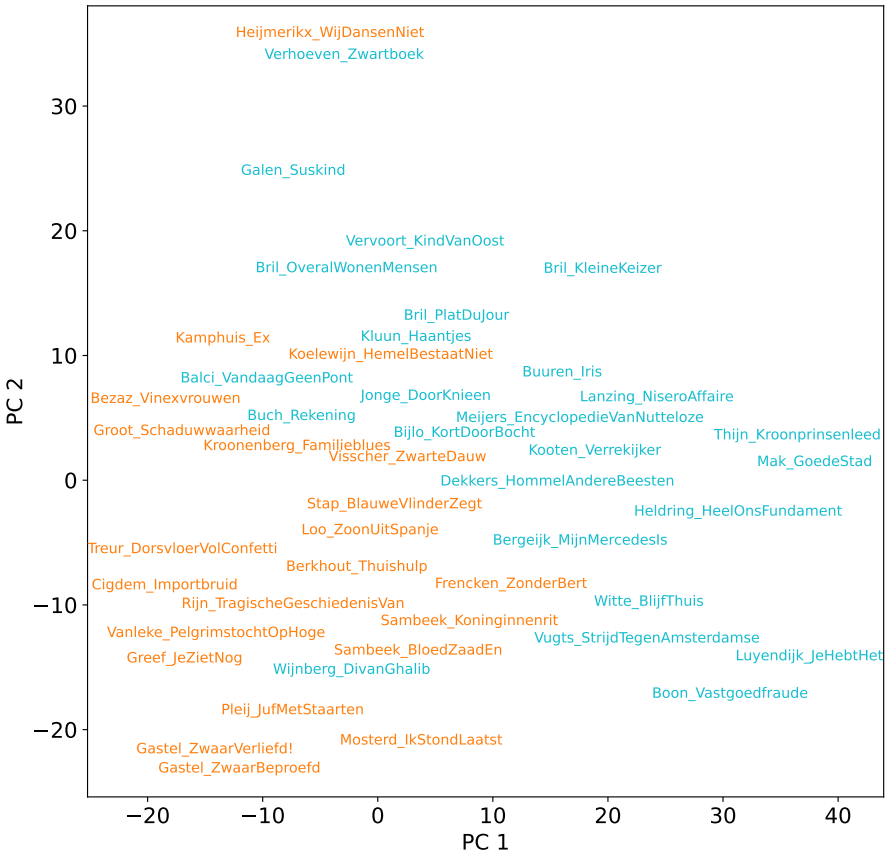

Figure 8: PCA showing the relative positions of the correctly classified not nominated books writing styles (NotNom). The books written by men are blue and the books written by women orange.

Zwagerman, Renate Dorrestein, and Vonne van der Meer.

As the Nom and the NomAut subset both consist of nominated authors, several authors occur in both Figure 4 and 5. For example, Herman Koch mostly positively correlates to misclassified Nom books, but negatively to misclassified NomAut books. Such a pattern is also seen in the correctly classified Nom and NomAut books by Harry Mulisch. Thus it seems that writing styles of Harry Mulisch and Herman Koch are more strongly related to nominated books than to not nominated books by nominated authors. This pattern could suggest that the writing style in nominated books is more related to the writing styles of Herman Koch and Harry Mulisch than the writing style seen in not nominated books by nominated authors. This is particularly interesting since Harry Mulisch is recognized as one of 'the three great' Dutch authors (*de grote drie*).

Figure 6–8 are Principal Component Analysis (PCA) plots, summarizing the similarity in writing styles for the correctly classified books. The books written by men are blue and the books written by women orange. Figure 6 shows that most correctly classified Nom books are clustered together in two closely related clusters, with a couple of outliers above or underneath the clusters. Interestingly, these outliers consist of works by Christiaan Weijts and Arnon Grunberg, which also positively correlate to the misclassified Nom books in the previously discussed heatmaps. The books written by women are placed in the two clusters, which shows that their writing style is closely related to the writing styles of nominated books written by men. Thus, the writing style of Nom books do not seem to differ based on author gender.

The NomAut books in Figure 7 are clustered more closely together, suggesting that the writing style in these books are more strongly related to each other. Figure 7 also shows that the writing style in NomAut books written by women does not strongly differ from the books written by men. The books by the authors that positively relate to the misclassified NomAut books are placed in the middle of the cluster. Interestingly, the works by Harry Mulisch are also centrally placed in the cluster.

In Figure 8 the books are clustered in a similar spread-out manner as in Figure 6. However, a stronger gender clustering is seen in the NotNom books than in the other classes, as the books written by women are predominantly placed on the left side. This could suggest that the writing style in NotNom books differs based on author gender.

In Figure 9–11, the relations between the correctly classified Nom, NomAut and NotNom books are shown in Bootstrap Consensus Trees (BCT) created with Stylo (Eder et al., 2016). The branches of the trees show which books are most similar to one another in terms of writing style. Compared to the previously shown heatmaps and PCA plots, a BCT is more robust, since it is based on clusterings of similar texts across different subsets of features, which ensures that only a consensus of robust similarities is visualized.

In the Nom and NomAut trees (Figure 9 and 10), multiple books of the same author are shown. For most authors, such as Christiaan Weijts (Nom) and Jeroen Brouwers (NomAut), these books are grouped together directly on the same branch. However, in each of the plots, not all books by the same author are grouped on the same branch, which would be expected given that the authorial signal typically dominates in stylometric analyses. For the Nom books, this concerns the books of Jeroen Brouwers and Arnon Grunberg, while for the NomAut books, this concerns the books of Harry Mulisch, Chaja Polak, J. Bernlef, and Joost Zwagerman, while for the NotNom books, this concerns the books of Martin Bril. Remarkably, *Elementen* by Harry Mulisch is placed on a separate branch in Figure 10.

In all three figures, distinctive main branches of writing styles can be identified, which are similar to the clustering shown in Figure 6–8. The left upper branch in Figure 11 consists predominately of works by women writers, suggesting that the distinction between writing styles is influenced by author gender.

Figure 9: Bootstrap Consensus Tree of correctly classified books, showing how the writing styles of nominated books (Nom) relate to each other.

Thus, no clear relation between writing style and nomination class or author gender could be identified, as the relation between the writing styles seems to depend highly on the distance between writing styles of particular authors. Several main branches of writing styles can be identified in each class, which do not show a clear division based on author gender except for the NotNom books. However, it is important not to draw general conclusions on gendered writing style, as the overlap in writing style between genders might not be reflected in the BCTs due to a strong influence of certain outliers (Koolen and van Cranenburgh, 2017). The results do show that the writing style of Harry Mulisch and Herman Koch are more closely related to nominated books than to not nominated books by nominated authors.

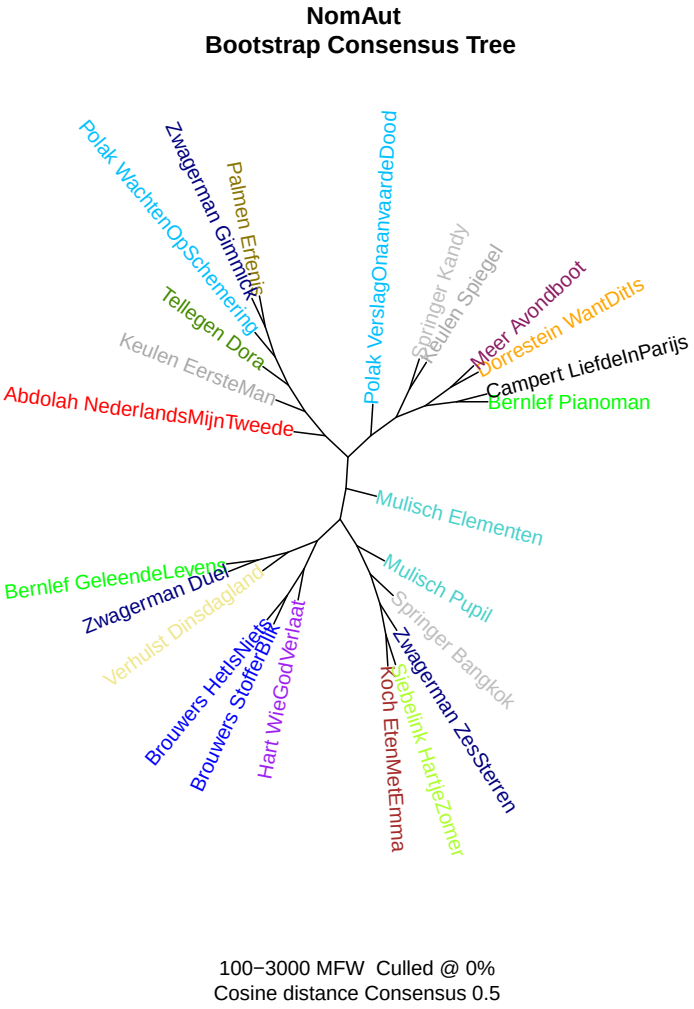

Figure 10: Bootstrap Consensus Tree of correctly classified books, showing related writing styles for not nominated books written by nominated authors (NomAut).

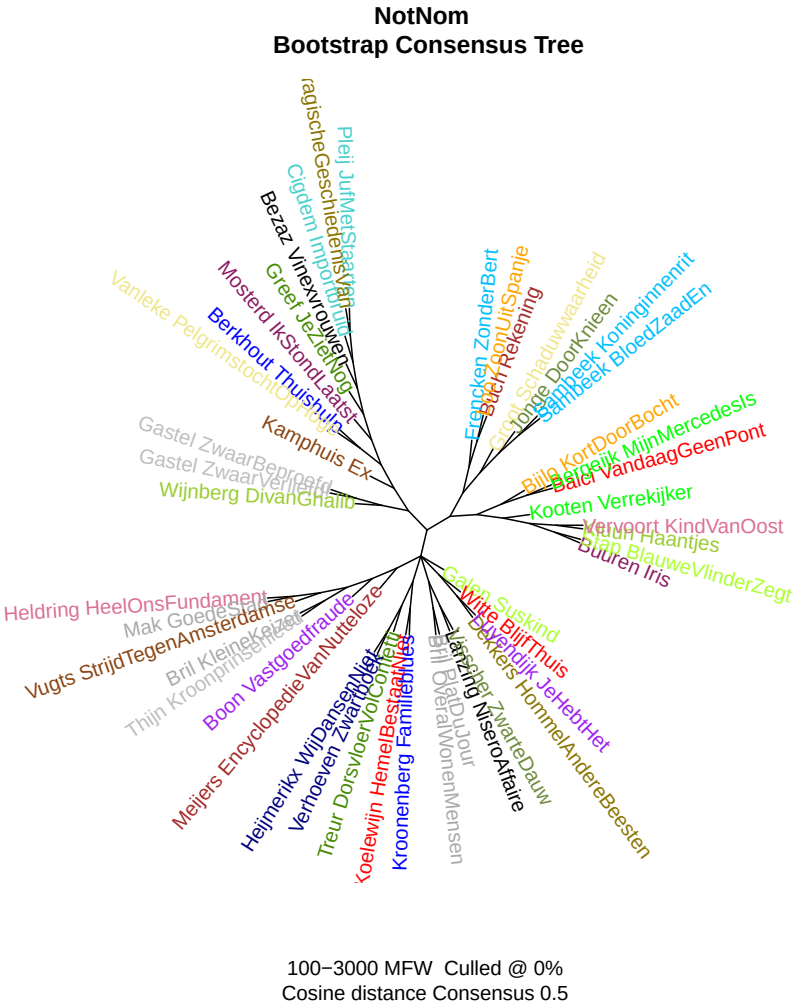

Figure 11: Bootstrap Consensus Tree of correctly classified books, showing related writing styles for not nominated books written by not nominated authors (NotNom).

5. Discussion

We have investigated author gender inequality in Dutch literary prizes using distant reading methods. We now address the three research questions presented in the Introduction (Section 1).

RQ1: To what extent can nominated and not nominated books be distinguished based on textual features alone? The results of the classification models show that it is possible to distinguish nominated and not nominated books to some extent based on textual features only. The three models on classification of nomination (nominated books (NOM), not nominated books from nominated authors (NOMAUT) and not nominated books from not nominated authors (NOTNOM), nominated-or-not and NOM or NOTNOM) all obtained an accuracy higher than chance. This means that the model predicts classes based on generalizations made on textual features. The results also show that the not nominated books by nominated authors (NOMAUT) are the hardest to classify, as this class consistently has the lowest F1-score (see Table 3) and accuracy scores (see Table 5). This could be due to the limited number of unique authors in this category, making it harder to generalize the distinguishing features for not nominated books by nominated authors. Another reason for the low performance of NOMAUT books is that this category is the least well defined. For example, the NOMAUT books also include *Boekenweekgeschenk* (book week gift) books , which are shorter books that are usually not awarded any literary prizes.

RQ2: Is there a relation between classifications of nominated versus not nominated books and author gender? The results of the classification tasks can be related to author gender. The results show a relation between the word use in books written by women and not nominated books by not nominated authors (NOTNOM). This relation is most clearly shown by the scores of the classification tasks. Books written by women consistently have the highest score for the NOTNOM class, in comparison to NOM and NOMAUT. This shows that for a classification task on nominated and not nominated books, it is easiest to classify NOTNOM for books written by women. For the classification task on author gender, NOTNOM has the highest scores on books written by women. Thus, there seems to be a relation between books written by women and NOTNOM.

For the books written by men, such a relation between the NOM, NOMAUT and NOTNOM classes was not found, but the books written by men did consistently have higher results than the books written by women, for all classes. This was probably not due to the higher number of books by men in the dataset, as this pattern was also seen in the subset with an equal author gender balance, although it was not as strong there.

It is important not to draw general conclusions on gendered word use based on these classifications, as the relation to author gender might be subtle and influenced by outliers. The overlap between books of authors of different genders can be larger than is portrayed in these results (Koolen, 2018). Furthermore, gender remains a social construct and gendered word use is strongly related to word use in gendered social groups (Bamman et al., 2014a; Butler, 1998; Nguyen et al., 2014; Oyewumi, 2002).

RQ3: Are the differences in topics/writing styles between books that are nominated for literary prizes and those that are not, related to author gender? For the topics, a few topics could be identified that occur relatively more in nominated or not nominated

books, which could sometimes be related to author gender. For example, the topic 'Second World War' is more common in nominated books and the topic 'family' is more common in not nominated books by nominated authors. Both topics have a high probability to occur in books written by men and books by women, respectively.

Other topics seem to relate to a specific nomination class but are actually more gender specific. The topic 'driving a car' is relatively more common in NomAut books, in comparison to Nom and NotNom books. When specifying the topics by author gender, the topic actually relates to books written by men in all three classes, but most strongly to NotNom books written by men. Another interesting result is that some topics seem to be judged to be of higher literary quality when the book was written by a man, as they occur specifically in nominated books written by men. For example, the topic 'English words' occurs most in Nom books written by men and not nominated books (both NomAut and NotNom) written by women. This supports the theory that for particular topics and genres, the judgment of literary quality of particular topics or genres is higher when a book is written by a man (Koolen et al., 2020).

For the difference in writing styles between nominated and not nominated, it was expected that the relation between books that are consistently correctly classified in the logistic regression models could be related to author gender. Such a pattern could neither be identified nor falsified. Another expectation was that a pattern could be found in the relation between correctly classified books and misclassified books. For the books by nominated authors (Nom and NotNom) such a pattern cannot be seen. The relation between the writing styles seems to highly depend on how close the writing styles of particular authors are related to each other. The results show that the writing style of Harry Mulisch and Herman Koch is more closely related to nominated books than to not nominated books by nominated authors. This could suggest that a particular writing style in Dutch literature exists which is more often nominated for literary prizes and which is more closely related to the writing styles of Harry Mulisch and Herman Koch.

6. Conclusion and Future Work

We have shown that author gender inequality in Dutch literary prizes can be investigated using distant reading methods. In addition, our results support the notion that the inequality in Dutch literary prizes is rooted in a homogeneous writing style that is related to the writing style of men. The predictive modeling results show that nominated and not nominated books are distinguishable, both for men and women writers, thus indicating that the nomination for literary prizes and literary quality is associated with particular word use. However, this word use seems to be further removed from women writers, particularly from their word use in nominated books, as the classification of books written by women consistently has the lowest performance. The analysis of the topics in nominated and not nominated books indicate that the relation between nominated and not nominated books and author gender is rather complex, and depends on the topic which is investigated. The difference in writing style of nominated and not nominated books cannot be clearly defined, but the results do suggest that there is a similarity between the writing style of Harry Mulisch and Herman Koch and writing styles that are nominated for literary prizes.

Future work The conclusions of this research are limited by several factors. Firstly, the dataset, in particular the number of unique authors, is rather limited. As authors have

a very distinguishable personal writing style (Herrmann et al., 2021; Tuzzi and Cortelazzo, 2018), having more authors in the dataset would lead to more different writing styles in the corpus, and therefore more generalizable results. Another improvement of the dataset would be to select not nominated books which were sent in by their publishers but were not selected for the long list. In this manner, the actual opponents of the nominated books could be used, as the goal is to select not nominated books that in theory could have been opponents of the nominated books.

Secondly, this research only focuses on author gender inequality, in particular between men and women, without taking into account other factors influencing language use, such as ethnicity, age and social class (Eckert, 2012). In order to analyze inequality in Dutch literary prizes, this should be researched as well, preferably in an intersectional manner. Additionally, influences within the literary environment, such as the prestige of a publisher or the reviews of books, could be considered as well. It would also be interesting to further research the textual factors that relate to the author gender inequality that have been identified. For example, the extent to which the writing styles of highly acclaimed writers, such as Harry Mulisch, relate to the general writing style of nominated books could be an avenue of future research.

Lastly, we would like to further investigate the potential relation of author gender with textual features, using additional classification experiments with nominated and not nominated books. One approach is to train only on books written by men, and see how this affects the classification score, and vice versa for books written by women. Differences with the results presented in this article could give more insight into whether or not the classifier picks up on gender bias through patterns in word usage it is trained on.

The combination of techniques used in this paper could also be applied to research other (potential) forms of inequality in the Dutch literary scene, such as ethnic and cultural background, socio-economic class and queerness. These techniques could also be used outside of the literary scene, for example to research inequality in job applications or grading bias in education.

Acknowledgments

We are grateful to DBNL for providing part of the dataset used in this research, and to Dong Nguyen and the anonymous reviewers for their feedback.

References

Karin Amatmoekrim. Een monoculturele uitwas: De ondraaglijke witheid van de Nederlandse letteren. *De Groene Amsterdammer*, August 2015.

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. Gender, genre, and writing style in formal written texts. *Text*, 23(3):321–346, 2003. ISSN 18607349. doi: 10.1515/text.2003.014.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 2014a. ISSN 14679841. doi: 10.1111/josl.12080.

David Bamman, Ted Underwood, and Noah A. Smith. A bayesian mixed effects model of literary character. *52nd Annual Meeting of the Association for Computational*

Linguistics, ACL 2014 - Proceedings of the Conference, 1:370–379, 2014b. doi: 10.3115/v1/p14-1035.

Pauwke Berkers. *Classification into the literary mainstream? Ethnic boundaries in the literary fields of the United States, the Netherlands and Germany, 1955-2005*. Erasmus University Rotterdam, 2009.

Douglas Biber and Edward Finegan. Styles of stance in english: Lexical and grammatical marking of evidentiality and affect. *Text-interdisciplinary journal for the study of discourse*, 9(1):93–124, 1989.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

Petra Boudewijn. 'and the award goes to...' women on the dutch literary award scene. *Journal of Dutch Literature*, 11(1), 2020. URL https://www.journalofdutchliterature.org/index.php/jdl/article/view/198.

John Burrows. 'delta': a measure of stylistic difference and a guide to likely authorship. *Literary and linguistic computing*, 17(3):267–287, 2002.

Judith Butler. *Imitation and Gender Insubordination*, pages 722—-730. Blackwell Publishers, 1998.

Jeroen Dera. The cultural diversity of text selection in Dutch literary education: An analysis of reading tips, teaching packs, and student choices., 2020. preprint on webpage at https://doi.org/10.31234/osf.io/a9tuq.

Jeroen Dera. De helaasheid der leeslijsten. over diversiteit in het literatuuronderwijs. *De Lage Landen*, 64 (1):115–121, 2021.

Margot Dijkgraaf and René Appel. *Vrouwen, mannen en de Libris Literatuur Prijs*. Stichting Libris Literatuur Prijs, 2013.

Penelope Eckert. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology*, 41:87–100, 2012. ISSN 00846570. doi: 10.1146/annurev-anthro-092611-145828.

Maciej Eder, Jan Rybicki, and Mike Kestemont. Stylometry with r: a package for computational text analysis. *R Journal*, 8(1):107–121, 2016. URL https://journal.r-project.org/archive/2016/RJ-2016-007/index.html.

Stefan Evert, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt. Understanding and explaining delta measures for authorship attribution. *Digital Scholarship in the Humanities*, 32(suppl_2):ii4–ii16, 2017.

Ethan Fast, Tina Vachovsky, and Michael S. Bernstein. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. *Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016*, pages 112–120, 2016.

Fixdit. Over ons, 2023. URL https://fixdit.nu/over-ons/.

Andrew Goldstone and Ted Underwood. The quiet transformations of literary studies: What thirteen thousand scholars could tell us. *New Literary History*, 45(3):359–384, 2014.

Susan C. Herring and John C. Paolillo. Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4):439–459, 2006. ISSN 13606441. doi: 10.1111/j.1467-9841.2006.00287.x. URL https://doi.org/10.1111/j.1467-9841.2006.00287.x.

J Berenike Herrmann, Arthur M Jacobs, and Andrew Piper. Computational stylistics. *Handbook of Empirical Literary Studies*, page 451, 2021.

Kim Jautze, Andreas van Cranenburgh, and Corina Koolen. Topic modeling literary quality. In *Digital Humanities 2016: Conference Abstracts*, pages 233–237, Krákow, Poland, 2016. URL http://dh2016.adho.org/abstracts/95.

Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211, 2019.

Os Keyes, Chandler May, and Annabelle Carrell. You keep using that word: Ways of thinking about gender in computing research. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–23, 2021.

Corina Koolen. *Reading beyond the female*. PhD thesis, University of Amsterdam, 2018. URL https://pure.uva.nl/ws/files/23823454/Thesis.pdf.

Corina Koolen and Andreas van Cranenburgh. These are not the Stereotypes You are Looking For: Bias and Fairness in Authorial Gender Attribution. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 12–22, 2017.

Corina Koolen, Karina van Dalen-Oskam, Andreas van Cranenburgh, and Erica Nagelhout. Literary quality in the eye of the Dutch reader: The National Reader Survey. *Poetics*, 79(February):101439, 2020. ISSN 0304422X. doi: 10.1016/j.poetic.2020.101439. URL https://doi.org/10.1016/j.poetic.2020.101439.

Timo Koren and Christine Delhaye. Depoliticising literature, politicising diversity: ethno-racial boundaries in Dutch literary professionals' aesthetic repertoires. *Identities*, 26(2):184–202, 2019. ISSN 15473384. doi: 10.1080/1070289X.2017.1391561. URL https://doi.org/10.1080/1070289X.2017.1391561.

Gong Lejun, Tang Xiangyu, and Li Huakang. Analysis of literary based on deep emotional network. In *2021 7th International Conference on Big Data Computing and Communications (BigCom)*, pages 227–233. IEEE, 2021.

Maksym Lupei, Alexander Mitsa, Volodymyr Repariuk, and Vasyl Sharkan. Identification of authorship of ukrainian-language texts of journalistic style using neural networks. *Eastern-European Journal of Enterprise Technologies*, 1(2):30–36, 2020. doi: 10.15587/1729-4061.2020.195041.

John Marsden, David Budden, Hugh Craig, and Pablo Moscato. Language individuation and marker words: Shakespeare and his maxwell's demon. *PloS one*, 8(6): e66813, 2013.

Dong Nguyen, Dolf Trieschnigg, A. Seza Dŏgrüoz, Rilana Gravel, Marïet Theune, Theo Meder, and Franciska De Jong. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. *COLING 2014 - 25th International Conference on Computational Linguistics, Proceedings of COLING 2014: Technical Papers*, pages 1950–1961, 2014.

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. Computational sociolinguistics: A survey. *Computational Linguistics*, 2016. ISSN 15309312. doi: 10.1162/COLI_a_00258.

Dong-Phuong Nguyen. *Text as social and cultural data: a computational perspective on variation in text*. PhD thesis, University of Twente, March 2017. SIKS dissertation series no. 2017-09.

Oyeronke Oyewumi. Conceptualizing gender: the eurocentric foundations of feminist concepts and the challenge of African epistemologies. *Jenda: A Journal of Culture and African Women Studies*, 2(1):1–9, 2002.

Anil Ramdas. Moedwil en kwade trouw bij blanke schrijvers. Niemand heeft oog voor het vreemde. *NRC Handelsblad*, March 1997.

Ebissé Rouw. Literatuur blijft te wit. *NRC Handelsblad*, May 2015.

Benjamin M Schmidt. Words alone: Dismantling topic models in the humanities. *Journal of Digital Humanities*, 2(1):49–65, 2012.

Roel Smeets, Eric Sanders, and Antal van den Bosch. Character centrality in present-day dutch literary fiction. *Digital Humanities Benelux Journal*, 2019.

Jean-François Staszak. Other/otherness. In *International Encyclopedia of Human Geography*. Elsevier Science, 2009.

Arjuna Tuzzi and Michele A Cortelazzo. What is elena ferrante? a comparative analysis of a secretive bestselling italian writer. *Digital Scholarship in the Humanities*, 33(3): 685–702, 2018.

Ted Underwood, David Bamman, and Sabrina Lee. The Transformation of Gender in English-Language Fiction. *Journal of Cultural Analytics*, pages 1–25, 2018. doi: 10.22148/16.019.

Andreas van Cranenburgh and Rens Bod. A Data-Oriented Model of Literary Language. *arXiv preprint arXiv:1701.03329*, 2017.

Andreas van Cranenburgh and Corina Koolen. Results of a single blind literary taste test with short anonymized novel fragment s. In *Proceedings of LaTeCH-CLfL*, pages 121–126, 2020. URL https://www.aclweb.org/anthology/2020.latechclfl-1.14.

Lucas van der Deijl, Saskia Pieterse, Marion Prinse, and Roel Smeets. Mapping the Demographic Landscape of Characters in Recent Dutch Prose Journal of Dutch Literature. *Journal of Dutch Literature*, 7(1):20–42, 2016. URL http://www.revisor.nl/entry/2095/slachtoffers-positiebepaling.

Lucas van der Deijl, Antal van den Bosch, and Roel Smeets. The canon of Dutch literature according to Google. *Journal of Cultural Analytics*, 4(2):11046, 2019.

Paulo Varela, Edson Justino, Alceu Britto, and Flávio Bortolozzi. A computational approach for authorship attribution of literary texts using sintatic features. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 4835–4842. IEEE, 2016.

Niña Weijers. Vrouwen *schrijven* niet met hun *tieten*. *NRC Handelsblad*, November 2014.

A. Appendix: Corpus

Author	Author gender	Title	Year	Libris Lit. Prijs	Boekenbon Lit. prijs	Target	Balanced author gender subset
Thomas van Aalten	Man	De Schuldigen	2012	Longlist		Nom	Yes
Kader Abdolah	Man	De Boodschapper	2008			NomAut	Yes
Kader Abdolah	Man	De Koran	2008			NomAut	No
Kader Abdolah	Man	De Kraai	2011			NomAut	Yes
Kader Abdolah	Man	Het Nederlands Als Mijn Tweede Vaderland	1996			NomAut	No
Kader Abdolah	Man	Het Huis van de Moskee	2006	Longlist		Nom	Yes
Ayaan Hirsi Ali	Woman	De Zoontjesfabriek	2002			NotNom	Yes
Erdal Balci	Man	Vandaag Geen Pont	2009			NotNom	Yes
Kees van Beijnum	Man	Een Soort Familie	2010		Shortlist	Nom	No
Kees van Beijnum	Man	De Oesters van Nam Kee	2000			NomAut	No
Kees van Beijnum	Man	De Ordening	1998			NomAut	No
Abdelkader Benali	Man	De Stem van mijn Moeder	2010	Longlist		Nom	Yes
Marinus van den Berg	Man	Nooit te Oud	2007			NotNom	No
Jeroen Bergeijk	Man	Mijn Mercedes is niet te koop	2006			NotNom	Yes
Jet Berkhout	Woman	De Thuishulp	2009			NotNom	Yes
J. Bernlef	Man	Geleendelevens	2010			NomAut	Yes
J. Bernlef	Man	De Pianoman	2008			NomAut	Yes
J. Bernlef	Man	Buiten is het Maandag	2004	Shortlist	Shortlist	Nom	Yes
J. Bernlef	Man	Zijn Dood	2011			NomAut	Yes
Hanna Bervoets	Woman	Lieve Celine	2011			NomAut	Yes
Naima el Bezaz	Woman	Vinex Vrouwen	2010			NotNom	Yes
Vincent Bijlo	Man	Kort door de Bocht	2008			NotNom	Yes
Aliefka Bijlsma	Woman	Mede Namens Mijn Vrouw	2010			NotNom	Yes
Oscar van den Boogaard	Man	Majesteit	2010			NomAut	Yes
Oscar van den Boogaard	Man	Meer dan een Minnaar	2010		Shortlist	Nom	Yes
Vasco van der Boon	Man	De Vastgoedfraude	2009			NotNom	No
Johan de Boose	Man	De Poppenspeler en de Duivelin	2009			NomAut	Yes
Martin Bossenbroek	Man	De Boerenoorlog	2013		Shortlist	Nom	Yes
Désanne van Brederode	Woman	Stille Zaterdag	2011			NomAut	Yes
Désanne van Brederode	Woman	Door Mijn Schuld	2010	Longlist		Nom	Yes
Claudia de Breij	Woman	Dingen die fijn zijn	2009			NotNom	Yes
Martin Brester	Man	Hoi, leuk dat je mijn profiel bekijkt!	2009			NotNom	No
Stefan Brijs	Man	Post voor mevrouw Bromley	2012	Longlist		Nom	Yes
Stefan Brijs	Man	De Engelenmaker	2006			Nom	Yes
Martin Bril	Man	Overal Wonen Mensen	2011			NotNom	No
Martin Bril	Man	Vaarwel Evelien	2011			NotNom	Yes
Martin Bril	Man	De Kleine Keizer	2008			NotNom	Yes
Martin Bril	Man	Evelien 2 Gelukkig Niet	2003			NotNom	Yes
Martin Bril	Man	Plat du Jour	2011			NotNom	No
Jan Brokken	Man	De Wil en de Weg	2006			NomAut	Yes
Jan Brokken	Man	Zeedrift	2009			NomAut	Yes
Jeroen Brouwers	Man	Het is Niets	1993			NomAut	Yes
Jeroen Brouwers	Man	Bittere Bloemen	2011	Shortlist	Shortlist	Nom	Yes
Jeroen Brouwers	Man	Datumloze Dagen	2008	Shortlist		Nom	Yes
Jeroen Brouwers	Man	Stoffer & Blik	2004			NomAut	No
Herman Brusselmans	Man	De terugkeer van Bonanza	1995			NomAut	Yes
Herman Brusselmans	Man	Guggenheimer wast witter	1996			NomAut	No

Author	Author gender	Title	Year	Libris Lit. Prijs	Boekenbon Lit. prijs	Target	Balanced author gender subset
Herman Brusselmans	Man	Uitgeverij Guggenheimer	1999			NomAut	No
Herman Brusselmans	Man	Trager dan Snelheid	2010			NomAut	Yes
Herman Brusselmans	Man	Het Einde van Mensen in 1967	1999			NomAut	Yes
Miquel Bulnes	Man	Attaque	2007			NomAut	Yes
Maarten van Buuren	Man	Iris	2011			NotNom	No
Boudewijn Büch	Man	De Rekening	1990			NotNom	Yes
Remco Campert	Man	Dagboek van een Poes	2007			NomAut	Yes
Remco Campert	Man	De Scholier	2009			NomAut	Yes
Remco Campert	Man	Een Liefde in Parijs	2004			NomAut	Yes
Hülya Cigdem	Woman	Import Bruid	2008			NotNom	Yes
Eveline Crone	Woman	Het Puberende Brein	2008			NotNom	Yes
Luc Deflo	Man	Angst	2007			NotNom	Yes
Midas Dekkers	Man	De Hommel en Andere Beesten	2005			NotNom	Yes
Peter Delpeut	Man	Het Vergeten Seizoen	2008	Longlist		Nom	No
Bernard Dewulf	Man	Kleine Dagen	2010	Winner		Nom	No
Nico Dijkshoorn	Man	Nooit Ziek Geweest	2012			NotNom	No
Adriaan van Dis	Man	Tikkop	2011	Shortlist		Nom	No
Adriaan van Dis	Man	Leeftocht	2007			NomAut	Yes
Adriaan van Dis	Man	Een Barbaar in China	1987			NomAut	No
Adriaan van Dis	Man	De Wandelaar	2007			NomAut	Yes
Renate Dorrestein	Woman	Een Sterke Man	1995	Shortlist		Nom	Yes
Renate Dorrestein	Woman	De Leesclub	2010			NomAut	Yes
Renate Dorrestein	Woman	De Stiefmoeder	2011			NomAut	Yes
Renate Dorrestein	Woman	Echt Sexy	2007			NomAut	Yes
Renate Dorrestein	Woman	Een Hart van Steen	1998			NomAut	Yes
Renate Dorrestein	Woman	Heden Ik	1993			NomAut	Yes
Renate Dorrestein	Woman	Het Duister Dat Ons Scheidt	2003			NomAut	Yes
Renate Dorrestein	Woman	Het Hemelse Gerecht	1991			NomAut	Yes
Renate Dorrestein	Woman	Is Er Hoop	2013			NomAut	Yes
Renate Dorrestein	Woman	Mijn zoon heeft een sexleven en ik lees mijn moeder Roodkapje voor	2006			NomAut	Yes
Renate Dorrestein	Woman	Zonder Genade	2002		Shortlist	Nom	Yes
Renate Dorrestein	Woman	Ontaarde Moeders	1992			NomAut	Yes
Renate Dorrestein	Woman	Noorderzon	2009			NomAut	Yes
Renate Dorrestein	Woman	Want dit is mijn lichaam	1997			NomAut	Yes
Renate Dorrestein	Woman	Zolang er leven is	2015			NomAut	Yes
Renate Dorrestein	Woman	Voor liefde druk op f	1999			NomAut	Yes
Dirk Draulans	Man	Beagledagboek	2010			NotNom	Yes
Jessica Durlacher	Woman	Held	2010			NotNom	Yes
G.L. Durlacher	Man	Godvergeten Tijd	2009			NomAut	Yes
Anna Enquist	Woman	Contrapunt	2009	Shortlist		Nom	Yes
Anna Enquist	Woman	Het Geheim	1997			NomAut	Yes
Anna Enquist	Woman	Het Meesterstuk	1994			NomAut	Yes
Anna Enquist	Woman	Mei	2007			NomAut	Yes
Anna Enquist	Woman	De Verdovers	2012	Longlist		Nom	Yes
Anna Enquist	Woman	De Thuiskomst	2006	Longlist		Nom	Yes
Rob van Essen	Man	Alles komt goed	2013	Longlist		Nom	No
Louis Ferron	Man	Karelische Nachten	1990		Winner	Nom	Yes
Herman Franke	Man	Zoek op Liefde	2009	Longlist		Nom	Yes
Mylou Frencken	Woman	Zonder Bert	2009			NotNom	Yes
Louise O. Fresco	Woman	De Utopisten	2008	Shortlist		Nom	Yes
Alex van Galen	Man	Süskind	2012			NotNom	Yes
Rodaan Al Galidi	Man	De Autist en de Postduif	2009			NotNom	Yes
Chantal van Gastel	Woman	Zwaar Verliefd!	2008			NotNom	Yes
Chantal van Gastel	Woman	Zwaar Beproefd	2009			NotNom	Yes
Esther Gerritsen	Woman	Superduif	2011	Shortlist		Nom	Yes
Esther Gerritsen	Woman	Dorst	2013	Shortlist		Nom	Yes
Wim Gijsen	Man	Kring van Stenen	1989			NotNom	No
Wim Gijsen	Man	Groene Eiland	1990			NotNom	Yes

Author	Author gender	Title	Year	Libris Lit. Prijs	Boekenbon Lit. prijs	Target	Balanced author gender subset
Wouter Godijn	Man	De dood van een auteur die een beetje op Wouter Godijn lijkt	2008	Longlist		Nom	No
Anne-Gine Goemans	Woman	Glijvlucht	2012	Longlist		Nom	Yes
Anne-Gine Goemans	Woman	Ziekzoekers	2008	Longlist		Nom	Yes
Saskia Goldschmidt	Woman	De Hormoonfabriek	2013	Longlist		Nom	Yes
Renske Greef	Woman	En je ziet nog eens wat	2009			NotNom	Yes
Karin de Groot	Woman	Schaduwwaarheid	2011			NotNom	Yes
Arnon Grunberg	Man	De Joodse Messias	2005	Longlist	Shortlist	Nom	No
Arnon Grunberg	Man	De Asielzoeker	2004			Nom	No
Arnon Grunberg	Man	Huid en Haar	2011	Shortlist	Shortlist	Nom	Yes
Arnon Grunberg	Man	Fantoompijn	2000		Winner	Nom	No
Arnon Grunberg	Man	Onze Oom	2009	Shortlist		Nom	No
Kees 't Hart	Man	Hotel Vertigo	2013	Longlist		Nom	Yes
Kees 't Hart	Man	Ter Navolging	2004	Longlist	Shortlist	Nom	Yes
Maarten 't Hart	Man	Wie God verlaat heeft niets te vrezen: de Schrift betwist	2011			NomAut	Yes
Mariëtte Haveman	Woman	De Vrouwenvanger	2011	Longlist		Nom	Yes
Detlev van Heest	Man	De verzopen katten en de Hollander	2011	Longlist		Nom	Yes
A.F.Th. van der Heijden	Man	Het Schervengericht	2007	Longlist	Winner	Nom	Yes
A.F.Th. van der Heijden	Man	Weerborstels	1992			NomAut	Yes
A.F.Th. van der Heijden	Man	Tonio	2012	Winner		Nom	No
Ellen Heijmerikx	Woman	Blinde Wereld	2009			NotNom	Yes
Ellen Heijmerikx	Woman	Wij Dansen Niet	2011			NotNom	Yes
J.L. Heldring	Man	Heel ons fundament kraakt en andere kanttekeningen	2003			NotNom	Yes
Kristien Hemmerechts	Woman	In het land van Dutroux	2008	Longlist		Nom	Yes
Kristien Hemmerechts	Woman	Wit Zand	1993			NomAut	Yes
Kristien Hemmerechts	Woman	Een jaar als (g)een ander	2003			NomAut	Yes
Kristien Hemmerechts	Woman	De waar gebeurde geschiedenis van Victor en Clara Rooze	2005			NomAut	Yes
Kristien Hemmerechts	Woman	Donderdagmiddag Halfvier	2002			NomAut	Yes
Kristien Hemmerechts	Woman	Ann	2008			NomAut	Yes
Kristien Hemmerechts	Woman	Als een kinderhemd	2006			NomAut	Yes
Kristien Hemmerechts	Woman	De laatste keer	2004			NomAut	Yes
Joke Hermsen	Woman	De liefde dus	2009	Longlist		Nom	Yes
Marijke Hilhorst	Woman	De vader, de moeder en de tijd	2008			NotNom	Yes
Oek de Jong	Man	Pier en Oceaan	2013	Shortlist		Nom	No
Freek de Jonge	Man	Door de knieën	2004			NotNom	Yes
Atte Jongstra	Man	De avonturen van Henry II Fix	2008	Longlist		Nom	Yes
Lieve Joris	Woman	Zangeres op Zanzibar en andere reisverhalen	2008			NotNom	Yes
Lieve Joris	Woman	De Golf	2007			NotNom	Yes
Martine Kamphuis	Woman	Vrij	2011			NotNom	Yes
Martine Kamphuis	Woman	Ex	2011			NotNom	Yes
Marie Kessels	Woman	Ruw	2010	Shortlist		Nom	Yes
Frank Ketelaar	Man	Avond aan avond	2006			NotNom	No
Mensje van Keulen	Woman	Liefde heeft geen hersens	2012	Longlist	Shortlist	Nom	Yes
Mensje van Keulen	Woman	De Spiegel	2008			NomAut	Yes
Mensje van Keulen	Woman	De eerste man	2011			NomAut	Yes
Mensje van Keulen	Woman	Een goed verhaal	2010	Shortlist		Nom	Yes
Yvonne Keuls	Woman	Alles went behalve een vent	2009			NotNom	Yes
Geert Kimpen	Man	Rachel	2011			NotNom	Yes

Author	Author gender	Title	Year	Libris Lit. Prijs	Boekenbon Lit. prijs	Target	Balanced author gender subset
Kluun	Man	Komt een vrouw bij de dokter	2009			NotNom	Yes
Kluun	Man	Haantjes	2010			NotNom	Yes
Nathalie Koch	Woman	Streken	2007	Longlist		Nom	Yes
Herman Koch	Man	Denken aan Bruce Kennedy	2005			NomAut	Yes
Herman Koch	Man	Eten met Emma	2000			NomAut	Yes
Herman Koch	Man	Odessa Star	2003			NomAut	Yes
Herman Koch	Man	Het Diner	2010	Longlist		Nom	Yes
Herman Koch	Man	Zomerhuis met Zwembad	2012	Longlist		Nom	Yes
Herman Koch	Man	Red ons Maria Montanelli	1989			NomAut	No
Jannetje Koelewijn	Woman	De hemel bestaat niet	2011			NotNom	Yes
Kees van Kooten	Man	De Verrekijker	2013			NotNom	Yes
Yvonne Kroonenberg	Woman	Familieblues	2012			NotNom	Yes
Ernest van der Kwast	Man	Mama Tandoori	2010			NotNom	No
Tom Lanoye	Man	Sprakeloos	2010	Shortlist	Shortlist	Nom	No
Fred Lanzing	Man	De Nisero-affaire	2009			NotNom	No
Rik Launspach	Man	1953	2009			NotNom	Yes
Stan Lauryssens	Man	Rode Rozen	2004			NotNom	Yes
Joke van Leeuwen	Woman	Alles Nieuw	2009	Longlist	Shortlist	Nom	Yes
Joke van Leeuwen	Woman	Feest van het begin	2013		Winner	Nom	Yes
Tomas Lieske	Man	Dünya	2009			Nom	Yes
Celine Linssen	Woman	Duet	2007			NotNom	Yes
Tessa de Loo	Woman	Zoon uit Spanje	2004			NotNom	Yes
Karel Glastra van Loon	Man	De Onzichtbaren	2013			NomAut	Yes
Karel Glastra van Loon	Man	Lisa's adem	2000			NomAut	No
Karel Glastra van Loon	Man	De passievrucht	1999		Winner	Nom	Yes
Joris Luyendijk	Man	Je hebt het niet van mij, maar	2010			NotNom	Yes
Geert Mak	Man	De goede stad	2007			NotNom	Yes
Geert Mak	Man	Reizen zonder John	2012			NotNom	Yes
Vonne van der Meer	Woman	De vrouw met de sleutel	2012	Longlist		Nom	Yes
Vonne van der Meer	Woman	Eilandgasten	1999			NomAut	Yes
Vonne van der Meer	Woman	Laatste seizoen	2002			NomAut	Yes
Vonne van der Meer	Woman	De Avondboot	2001			NomAut	Yes
Vonne van der Meer	Woman	De reis naar het kind	1989			NomAut	Yes
Vonne van der Meer	Woman	Take 7	2007			NomAut	Yes
Hein Meijers	Man	Encyclopedie van nutteloze feiten	2012			NotNom	Yes
Doeschka Meijsing	Woman	Over de liefde	2008	Longlist	Winner	Nom	Yes
Jan van Mersbergen	Man	Naar de overkant van de nacht	2012	Longlist		Nom	No
Marente de Moor	Woman	De nederlandse maagd	2011			Nom	Yes
Margriet de Moor	Woman	Op de rug gezien	1989		Shortlist	Nom	Yes
Margriet de Moor	Woman	De schilder en het meisje	2011	Longlist		Nom	Yes
Maria Mosterd	Woman	Echte mannen eten geen kaas	2008			NotNom	Yes
Lucie Mosterd	Woman	Ik stond laatst voor een poppenkraam	2009			NotNom	Yes
Harry Mulisch	Man	De Pupil	1987			NomAut	Yes
Harry Mulisch	Man	De ontdekking van de hemel	1992		Shortlist	Nom	No
Harry Mulisch	Man	De Procedure	1999	Winner		Nom	Yes
Harry Mulisch	Man	De Elementen	1988			NomAut	Yes
Charlotte Mutsaers	Woman	Koetsier Herfst	2009	Shortlist		Nom	Yes
Marcel Möring	Man	Louteringsberg	2012	Longlist		Nom	Yes
Willem Nijholt	Man	Met bonzend hart	2011			NotNom	Yes
Nelleke Noordervliet	Woman	Zonder noorden komt niemand thuis	2010	Longlist		Nom	Yes
Nelleke Noordervliet	Woman	Vrij Man	2013	Longlist		Nom	Yes
Nelleke Noordervliet	Woman	Snijpunt	2009	Longlist		Nom	Yes
Michiel Klein Nulent	Man	Het Koekoeksei	2011			NotNom	No
Ellen Ombre	Woman	Maalstroom	1992			NotNom	Yes

Author	Author gender	Title	Year	Libris Lit. Prijs	Boekenbon Lit. prijs	Target	Balanced author gender subset
Connie Palmen	Woman	Logboek van een onbarmhartig jaar	2011			NomAut	Yes
Connie Palmen	Woman	De Wetten	1991			NomAut	Yes
Connie Palmen	Woman	De Erfenis	1999			NomAut	Yes
Connie Palmen	Woman	De Vriendschap	1995		Winner	Nom	Yes
Koen Peeters	Man	Grote Europese roman	2008	Shortlist		Nom	Yes
Rascha Peper	Woman	Vossenblond	2011			NomAut	Yes
Rascha Peper	Woman	Dooi	1999			NomAut	Yes
Rascha Peper	Woman	Een Spaans hondje	1998			NomAut	Yes
Rascha Peper	Woman	Wie scheep gaat	2003			NomAut	Yes
Yves Petry	Man	De maagd Marino	2011	Winner		Nom	Yes
Eefje Pleij	Woman	Juf met staarten krijgt een staartje	2008			NotNom	Yes
Chaja Polak	Woman	Verslag van een onaanvaarde dood	2007			NomAut	Yes
Chaja Polak	Woman	Wachten op de schemering	2007			NomAut	Yes
Anne Provoost	Woman	In de zon kijken	2008	Longlist		Nom	Yes
Anil Ramdas	Man	De papegaai, de stier, en de klimmende bougainvillea	1992			NotNom	Yes
David van Reybrouck	Man	Congo	2010	Winner		Nom	Yes
Elle van Rijn	Woman	De tragische geschiedenis van mijn succes	2006			NotNom	Yes
Thomas Rosenboom	Man	Zoete Mond	2010	Longlist		Nom	Yes
Thomas Rosenboom	Man	De nieuwe man	2003		Shortlist	Nom	Yes
Helga Ruebsamen	Woman	Beer is terug	2000	Shortlist		Nom	Yes
Ciel van Sambeek	Woman	Bloedzaaden	2011			NotNom	Yes
Ciel van Sambeek	Woman	Koninginnenrit	2008			NotNom	Yes
Peter Schaap	Man	De bruiden van Tyobar	1992			NotNom	Yes
Jaap Scholten	Man	De wet van Spengler	2009			NotNom	Yes
Jaap Scholten	Man	Morgenster	2009			NotNom	Yes
Jan Siebelink	Man	Verdwaald Gezin	1993			NomAut	No
Jan Siebelink	Man	Vera	1997			NomAut	Yes
Jan Siebelink	Man	Suezkade	2008			NomAut	Yes
Jan Siebelink	Man	Knielen op een bed violen	2005	Shortlist	Winner	Nom	No
Jan Siebelink	Man	De overkant van de rivier	1990			NomAut	Yes
Jan Siebelink	Man	Engelen van het duister	2001			NomAut	No
Jan Siebelink	Man	Hartje zomer	1991			NomAut	Yes
Jan Siebelink	Man	Het lichaam van Clara	2010			NomAut	Yes
Mart Smeets	Man	De Afrekening	2010			NotNom	Yes
Susan Smit	Woman	Wat er niet meer is	2007			NotNom	Yes
Susan Smit	Woman	Wijze Mannen	2010			NotNom	Yes
F. Springer	Man	Kandy	1998			NomAut	Yes
F. Springer	Man	Bangkok, een elegie	2005			NomAut	No
Rosalie Sprooten	Woman	De pest voor een schip	1989			NotNom	Yes
Sophie van der Stap	Woman	Een blauwe vlinder zegt gedag	2008			NotNom	Yes
Bianca Stigter	Woman	De ontsproten Picasso	2008		Shortlist	Nom	Yes
Henk van Straten	Man	Salvador	2012	Longlist		Nom	No
Henk van Straten	Man	Superlul	2011			NomAut	Yes
Henk van Straten	Man	Kleine Stinkerd	2008			NomAut	No
Toon Tellegen	Man	Dora	1998			NomAut	No
Peter Terrin	Man	Post mortem	2012	Longlist	Winner	Nom	No
Charles den Tex	Man	Cel	2009	Longlist		Nom	Yes
Charles den Tex	Man	Spijt	2009			NomAut	Yes
Charles den Tex	Man	De macht van meneer Miller	2005			NomAut	Yes
Christiaan Thijm	Man	Het proces van de eeuw	2011			NotNom	Yes
Ed van Thijn	Man	Kroonprinsenleed	2008			NotNom	Yes
P.F. Thomése	Man	De weldoener	2011		Shortlist	Nom	Yes
Anneloes Timmerije	Woman	Aus liebe	2009			NotNom	Yes

Author	Author gender	Title	Year	Libris Lit. Prijs	Boekenbon Lit. prijs	Target	Balanced author gender subset
Willem van Toorn	Man	Rooie en andere verhalen over mijzelf en mijn klas	1992			NotNom	No
Franca Treur	Woman	Dorsvloer vol confetti	2009			NotNom	Yes
Carolina Trujillo	Woman	De terugkeer van Lupe García	2009			Nom	Yes
Betsy Udink	Woman	Allah & Eva	2006			NotNom	Yes
Monica Vanleke	Woman	Pelgrimstocht op hoge hakken	2011			NotNom	Yes
Annelies Verbeke	Woman	Vissen redden	2010	Longlist		Nom	Yes
Alex Verburg	Man	Dwalingen	2009			NotNom	Yes
Paul Verhoeven	Man	Zwartboek	2006			NotNom	Yes
Dimitri Verhulst	Man	Godverdomse dagen op een godverdomse bol	2009			Nom	Yes
Dimitri Verhulst	Man	De helaasheid der dingen	2006	Longlist	Shortlist	Nom	No
Dimitri Verhulst	Man	Mevrouw Verona daalt de heuvel af	2007	Longlist	Shortlist	Nom	No
Dimitri Verhulst	Man	Problemski Hotel	2003			NomAut	No
Dimitri Verhulst	Man	De kamer hiernaast	1999			NomAut	Yes
Dimitri Verhulst	Man	Dinsdagland	2004			NomAut	Yes
Hans Vervoort	Man	Kind van de Oost	1992			NotNom	Yes
Hans Vervoort	Man	Geluk is voor de dommen	2003			NotNom	Yes
Rachel Visscher	Woman	Zwarte Dauw	2011			NotNom	Yes
Arjan Visser	Man	Paganinipark	2011			NomAut	Yes
Carolijn Visser	Woman	Vrouwen in den vreemde	2008			NotNom	Yes
Erik Vlaminck	Man	Brandlucht	2012	Longlist		Nom	No
Paul Vugts	Man	De strijd tegen de Amsterdamse onderwereld	2011			NotNom	Yes
Robert Vuijsje	Man	Alleen maar nette mensen	2009	Shortlist		Nom	Yes
Christiaan Weijts	Man	Via Cappello 23	2009		Shortlist	Nom	No
Christiaan Weijts	Man	Art 285b	2006			Nom	No
Christiaan Weijts	Man	De etaleur	2010	Longlist		Nom	No
Gerwin van der Werf	Man	Wild	2012	Longlist		Nom	Yes
Lodewijk Wiener	Man	De verering van Quirina T.	2007			Nom	No
Tommy Wieringa	Man	Caesarion	2009		Shortlist	Nom	No
Tommy Wieringa	Man	Dit zijn de namen	2013	Winner		Nom	No
Nachoem Wijnberg	Man	Politiek en liefde	2002			NotNom	Yes
Nachoem Wijnberg	Man	Divan van Ghalib	2009			NotNom	Yes
Leon de Winter	Man	Het recht op terugkeer	2008	Longlist	Shortlist	Nom	Yes
Patrick Witte	Man	Blijf Thuis	2009			NotNom	Yes
Ivan Wolffers	Man	Onweer in de verte	2009			NotNom	Yes
Annejet van der Zijl	Woman	Bernhard	2010			NotNom	Yes
Joost Zwagerman	Man	Transito	2007		Shortlist	Nom	Yes
Joost Zwagerman	Man	Duel	2011			NomAut	Yes
Joost Zwagerman	Man	Gimmick	1989			NomAut	No
Joost Zwagerman	Man	Vals licht	1992		Shortlist	Nom	Yes
Joost Zwagerman	Man	Zes sterren	2002			NomAut	Yes
Joost Zwagerman	Man	De buitenvrouw	2009			NomAut	No