Vector space explorations of literary language

van Cranenburgh, Andreas; van Dalen-Oskam, Karina; van Zundert, Joris

Published in:
Language Resources and Evaluation

DOI:
10.1007/s10579-018-09442-4

Publication date:
2019

Lang Resources & Evaluation (2019) 53:625–650
https://doi.org/10.1007/s10579-018-09442-4

ORIGINAL PAPER

Vector space explorations of literary language

Andreas van Cranenburgh[1] · **Karina van Dalen-Oskam**[2,3] · **Joris van Zundert**[2]

Published online: 9 February 2019

Abstract Literary novels are said to distinguish themselves from other novels through conventions associated with *literariness*. We investigate the task of predicting the literariness of novels as perceived by readers, based on a large reader survey of contemporary Dutch novels. Previous research showed that ratings of literariness are predictable from texts to a substantial extent using machine learning, suggesting that it may be possible to explain the consensus among readers on which novels are literary as a consensus on the kind of writing style that characterizes literature. Although we have not yet collected human judgments to establish the influence of writing style directly (we use a survey with judgments based on the titles of novels), we can try to analyze the behavior of machine learning models on particular text fragments as a proxy for human judgments. In order to explore aspects of the texts associated with literariness, we divide the texts of the novels in chunks of 2–3 pages and create vector space representations using topic models (Latent Dirichlet Allocation) and neural document embeddings (Distributed Bag-of-Words Paragraph Vectors). We analyze the semantic complexity of the novels using distance measures, supporting the notion that literariness can be partly explained as a deviation from the norm. Furthermore, we build predictive models and identify specific keywords and stylistic markers related to literariness. While genre plays a role, we find that the greater part of factors affecting judgments of literariness are

This work is part of The Riddle of Literary Quality, a project supported by the Royal Netherlands Academy of Arts and Sciences through the Computational Humanities program. In addition, the first author was supported by the German Research Foundation DFG.

✉ Andreas van Cranenburgh
a.w.van.cranenburgh@rug.nl

[1] Information Science, University of Groningen, Groningen, The Netherlands

[2] Huygens ING, Royal Netherlands Academy of Arts and Sciences, Amsterdam, The Netherlands

[3] Universiteit van Amsterdam, Amsterdam, The Netherlands

explicable in bag-of-words terms, even in short text fragments and among novels with higher literary ratings. The code and notebook used to produce the results in this paper are available at https://github.com/andreasvc/litvecspace.

Keywords Literature · Literariness · Document embeddings · Topic models

1 Introduction

Recent work has applied computational methods to the study of literary or general quality of prose (Louwerse et al. 2008; Ashok et al. 2013; Crosbie et al. 2013; Maharjan et al. 2017) and poetry (Underwood 2015). In particular, the task considered in this paper of predicting the literary prestige of Dutch novels has been addressed before (van Cranenburgh and Koolen 2015; van Cranenburgh and Bod 2017), as part of a project called The Riddle of Literary Quality.[1] It was shown that judgments of literariness, the degree to which a text is perceived as literary, can be predicted to a substantial extent using machine learning based on textual characteristics. Empirically, there is agreement among readers on the literariness of books, and the success of predictive models confirms that this agreement is reflected in the texts of the novels to a substantial degree. However, what is lacking is an explanation of the mechanisms by which text-intrinsic features contribute to the literary prestige of a text. In this paper we focus on investigating stylistic mechanisms; we use the following, broad definition of style:

> Style is a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively (Herrmann et al. 2015).

This paper uses the same Riddle data set but looks at smaller passages of text for two reasons: (a) to test intuitions on the nature of literariness and how it is reflected in texts, and (b) to get a better idea about the textual characteristics that influence the accuracy of predictions with particular computational methods.

Superficially, this task is a document classification task just like common NLP benchmarks with reviews of IMDB (Maas et al. 2011) and Yelp[2] that evaluate the prediction of ratings and sentiment polarity. However, compared to sentiment polarity, literariness is a much less transparent notion. Many words are strongly associated with a sentiment, while literariness can manifest itself in less concrete aspects such as complexity and layers. Although contextual factors such as negation and irony do complicate sentiment classification, words loaded with sentiment (great, terrible, exciting, etc.) are clear give-aways in a review, and the review itself has the direct goal to express its sentiment. Therefore we can expect that the classification of a review can be readily attributed to a limited set of surface features (words or phrases) that explain why a review is positive or negative (called a rationale in Lei et al. 2016).

[1] http://literaryquality.huygens.knaw.nl.

[2] https://www.yelp.com/dataset/.

There is a recent trend towards automating the explanation and interpretation of black box machine learning models (Ribeiro et al. 2016), or models that are interpretable by design, e.g. through attention mechanisms (Yang et al. 2016). However, these methods are limited to explaining their predictions in terms of the relative importance of features such as particular words. Explanations in terms of higher-level patterns will most likely continue to rely on manual application of domain knowledge.

These attempts at explaining the results of machine learning techniques, although highly exploratory still, should warrant the interest of digital humanities researchers and of their critics. The idea of natural language processing techniques as impenetrable black boxes is common among critics of the field of digital humanities (e.g., Fish 2012). The results derived from the application of such technologies is suspect because these black boxes handle vast amounts of data far beyond the ability of human interpretable aggregation but cannot be studied critically. Equally commonplace is the criticism that results of the application of machine learning techniques to problems of literary criticism are intellectually underwhelming (e.g., Brennan 2017). Straw man criticism can often be readily refuted (cf. Kirschenbaum 2014) but the relatively hermetic mathematical nature of many machine learning technologies presents problems of interpretation that digital humanities researchers themselves grapple with still (Clement et al. 2008; Sculley and Pasanek 2008). However, as Ted Underwood has argued, even if still hard to explain these methods allow us to understand the methodology of literary history as more than a zero-sum game of critical interpretation. That is: we can now query whole bodies of literature to evaluate if, for instance, indeed first person perspective is prevalent in psychological novels. And in contrast to what is often asserted the machine learning tools we can apply do accommodate the intentional blurriness of our literary categories and definitions (Underwood 2013). With the method under investigation here we aim to contribute to this broader methodological issue of literary research. But our aim is equally to be critical of the methods we apply. It is kind of a methodological myth-in-the-making that machine learning techniques are too complicated to understand and that it is all but impossible to explain how they yield the results they yield. However, attention mechanisms contribute to the progress in our abilities to explain how machine learning techniques arrive at their answers (cf. Kestemont and Stutzmann 2017)—and so does the critical interrogation of the results yielded by such methods as, for instance, exercised in this paper.

2 Corpus and survey data

A corpus of 401 recent Dutch novels was selected, consisting of the most popular books in 2010–2012. Popularity was based on figures of book sales and library loans. A large reader survey was conducted to collect data on the perceived literariness and quality of these novels.

The respondents were first asked to indicate which of the 401 books they had read. Then they were presented with a randomly selected list with the author and title of seven books they indicated to have read and were asked to rate these. Ratings

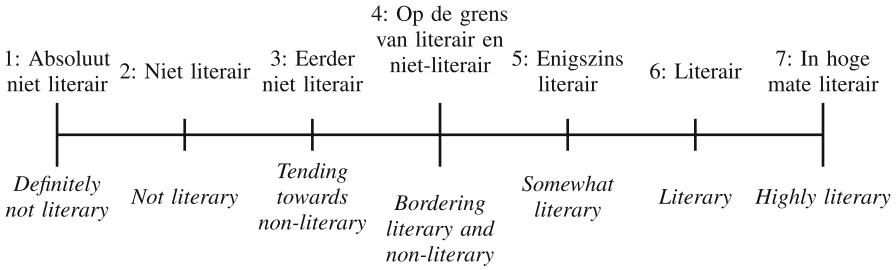

Fig. 1 The Likert scale used in the survey to collect literary ratings given the author and title of a novel. A similar scale was used for the quality ratings

were collected on a 1–7 Likert scale; see Fig. 1. Respondents could also answer "don't know"; these ratings are not used in this paper. If they wished, they could ask for another set of seven books to rate.

Since the ratings were provided with respect to the title and author of each novel, respondents were expected to provide their judgments from memory, without being presented with the text of the novels. They were also not provided with any definition of what literariness or literature is supposed to be, to encourage them to provide their own intuitions of what literariness is and not one provided by us. In addition to asking about novels that respondents had read, as a control, respondents were also asked to provide judgments for seven books they had not read.

About 14k respondents from the general public participated in the online survey. The online survey was open to everyone. In their motivation of their score for one of the books they rated, some respondents self-identified as being professional literary critics. While we could not guarantee that respondents could take the survey only once, inspecting the IP addresses and times of submission did not reveal suspicious activity. When the same IP address did occur in multiple submissions, the IP address was from an organization such as a library where multiple submissions are to be expected, and there were no patterns in the ratings which might indicate an attempt at manipulation.

We use the mean rating for each novel as ground truth in this work. Analysis of the variance of the ratings shows that for novels with at least 50 ratings there is consensus among the ratings: the t-distributed 95% confidence interval of the ratings for 91% of those novels has a width smaller than 0.5; e.g., given a mean of 3, the confidence interval typically lies within 2.75–3.25. Respondents were also asked to motivate some of their judgments; the answers indicate that writing style plays a role in their judgments, see Table 1.

In order to increase the number of data points and to zoom in on more specific aspects of the texts, we divide the texts of novels into chunks. We split the books in chunks of approximately 1000 tokens (i.e., 2–3 pages of text), rounded up or down to the nearest sentence boundary. The texts are converted to lower case and tokenized with the tokenizer of the Alpino parser (Bouma et al. 2001). See Table 2 for basic statistics of the corpus. The participants of the survey were asked to rate the novels as a whole and were not asked about specific features such as style or

Table 1 Some responses by respondents on the question: "Why did you rate this book with the score for literariness as you did?"

– The writing style
– Great, suspenseful and surprising book. Writing style not that surprising
– I did not like the writing style
– The book has a lot of depth and multiple layers
– It is suspenseful, the storyline is perfect, but in a literary novel I expect a deeper layer
– Shallow story, one-dimensional characters, no deeper layers

Table 2 Statistics for the corpus of novels

Novels	401
Chunks	52,107
Sentences	5,061,017
Tokens	52,320,029
Mean tokens per chunk (SD)	1000.2 (7.3)

narrative. Therefore each of the chunks from a novel is associated with the same rating. This is of course a compromise, because there are bound to be stylistic and narrative differences across the chunks of the books.

A proper investigation of this intra-textual variance in literariness would require a survey with human judgments on the level of text fragments, because such data is needed to directly establish the influence of textual features on human judgments. However, by studying the variance of predictive models on the level of text fragments, we can already get an idea of the kind of textual features that may be associated with different levels of literariness, even though confirmation would require an additional survey to be conducted.

3 Unsupervised document representations

The Vector Space Model of language assigns coordinates to documents in a high-dimensional space in which semantic similarity of documents and words is realized as spatial distance. We apply several unsupervised methods for creating such vector spaces from texts; unsupervised refers here to a model that is based strictly on the text of the novels, and is not trained to predict a specific variable such as the literary rating. In a later section we apply supervised predictive models that take the ratings of the novels into account and predict them from the document vectors.

3.1 Baseline: bag of words

An extremely simple yet strong document representation is the Bag-of-words (BoW) model. A bag is an unordered set in which each member is associated with a count. We use this model as a baseline. Each document is represented as a vector of word counts. We considered using term frequencies and *tf-idf*; in the end reducing

the word counts to binary features performed best. A limitation of this representation is that information on word order is discarded, and related words are represented as independent dimensions without exploiting their distributional properties. A simple way to retain a modicum of word order information is to count not just words (unigrams), but occurrences of two consecutive words (bigrams). We opt for words and not characters as the basic unit of analysis because words are more helpful when interpreting the results.

3.2 Topic modeling

Latent Dirichlet Allocation (LDA; Blei et al. 2003) is a Bayesian topic model that learns distributions of topics across words and documents. The input for LDA is preprocessed with lemmatization and the removal of function words and names. The output dimension (number of topics) is 50. We re-use the model presented in Jautze et al. (2016), which was obtained with Mallet (McCallum 2002).

LDA topic models are popular in Digital Humanities because individual topics can often be readily interpreted as coherent themes of related words. Since LDA applies a Dirichlet prior[3] to the topic and word distributions that it learns, there is a tendency for a small number of items to receive a large share of the weights, and a long tail of less relevant items. This helps interpretation because these prominent items (topics or words) stand out; this is in contrast to other models in which the weights are spread out over a large number of features, making interpretation more challenging.

3.3 Neural document embeddings

Paragraph Vectors (also referred to as doc2vec; Le and Mikolov 2014) are neural document embeddings based on an extension of word2vec to sequences of arbitrary length (the term paragraph should be taken loosely as a sequence that can be anything from sentences to documents). Compared to the aforementioned models, paragraph vectors have two advantages: they do not completely ignore word order in documents (by considering a small moving context window), and they learn more fine-grained aspects in which context affects meaning because words are not assigned to a fixed number of discrete topics. Compare this with the previous models that use a BoW representation as input, which represents a document as a list of word counts: while word co-occurrence in a document is represented, information on whether words tend to occur close together is lost, which would give more information on their relatedness.

We use the Distributed Bag-of-Words (DBoW) model with negative sampling as implemented in gensim (Řehůřek and Sojka 2010).[4] While Le and Mikolov (2014) reports that the Distributed Memory model is superior, later work such as Lau and

[3] Note that the Dirichlet prior here refers to a preference for a shape of the probability distributions, not the use of any prior information on prominent words or topics.

[4] We use the recently released version 3.5.0, which contains an important bug fix related to the learning rate.

Baldwin (2016) report that the simpler DBoW model is more effective. The mechanism by which DBoW paragraph vectors are trained is based on a pseudo-task of predicting the neighboring context words for each word in a paragraph. This task is learned by jointly optimizing two kinds of representations, those of words and of paragraphs. Training consists in changing the representation for a paragraph so as to maximize the number of words that can be correctly guessed given its representation. Aside from predicting words that occur in the context, negative predictions are made for unrelated words that do not occur in the paragraph (negative sampling). A correction is introduced to avoid highly frequent words being overrepresented during negative sampling.

We set the dimensions of the paragraph vectors to 300 and the window size to 10 words. Due to the more fine-grained semantics that paragraph vectors can represent, we choose a higher number of dimensions compared to the 50 topics described in the previous section. Apart from ignoring words with a count below 10, no further preprocessing is done on the tokenized texts (e.g., punctuation is kept). The model includes a word embedding model in the same vector space as the paragraph vectors, such that distances between word and paragraph vectors can be queried (Dai et al. 2015). The end result is a vector space with three important properties:

1. Words that commonly co-occur in similar contexts are close together; typically semantically related words.
2. Paragraphs with similar semantics are close together.
3. A paragraph and word are close together if they are semantically related.

4 Literariness as semantic complexity

A common suggestion is that literary novels are distinguished by being more creative, original, or unique with respect to other novels, which are said to more closely follow established genre tropes. A particularly strong statement of this claim is given by Louwerse (2004, p. 220):

> [...] the lack of internal homogeneity in one text, between texts and between authors can be explained by the (semantic) deviation from the norm the author tries to establish. These variations are exactly what makes the idiolect and sociolect of literary texts unique, and is in fact what makes those texts literary.

In this section we will operationalize and test this in several ways. We explore several ways in which textual distances (deviation) can be used to investigate the role of semantic complexity in literary novels compared to other novels.

Following the vector space model of language, the document vectors of the chunks of the novels can be interpreted as coordinates in vector space. In particular, Euclidean distance provides a geometric operationalization of contextual similarity among document vectors. We will consider both the vectors for chunks of novels, as well as vectors for the whole novel. In the latter case we take the centroid of its

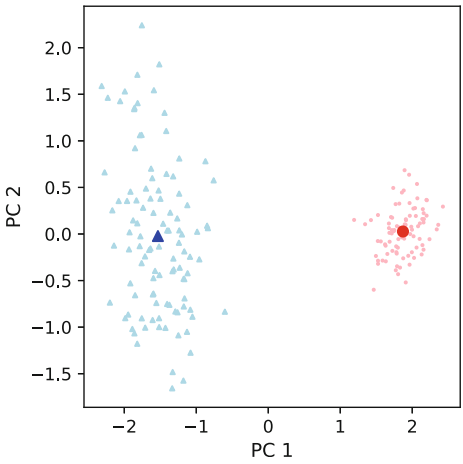

Fig. 2 An example of intra-textual variance, visualized with a PCA plot of the vectors of the novel chunks and their centroids. A larger distance between points represents a larger semantic variance. Left: Wieringa, *Caesarion* (high variance). Right: Slee, *Fatale Liefde* (Fatal Attraction; low variance)

chunks as representative for the whole novel; i.e., the mean vector across the chunks of a novel. Geometrically, the centroid is the center of gravity of a set of points.

4.1 Intra-textual variance

A simple measure of semantic complexity is the *intra-textual variance* of the document vectors of a novel. We can think of a novel as a cloud of points in vector space and this cloud can be either dense or expansive, depending on the semantic similarity of its chunks. Here dense refers to a set of chunks which are highly similar to each other, while expansive refers to a large variance in semantic similarity. We measure the semantic variance of a novel[5] by measuring the Euclidean distance of its chunks to the centroid of the novel. The result is summarized as the mean of squared distances:

$$\text{variance}(T) = 1/|T| \sum_{i=0}^{i<|T|} ||\mu_T - t_i||^2$$

where μ_T the centroid of the novel T and t_i is its ith chunk. This method of comparing elements to a centroid is the same heuristic used by the K-means algorithm to identify clusters. See Fig. 2 for an illustration. This visualization, as well as the ones that follow, is based on a Principal Components Analysis (PCA) dimensionality reduction.

4.2 Stepwise distance

A variation on this is to measure the variance between each pair of consecutive chunks, again summarized as the mean of squared distances; we call this the *stepwise distance*:

[5] Semantic variance can be seen as a more fine-grained version of the topic diversity presented in Jautze et al. (2016), which showed that genre-novels tend to concentrate on a single topic, while literary novels tend to be spread out over more topics. This version is not restricted to a fixed number of discrete topics.

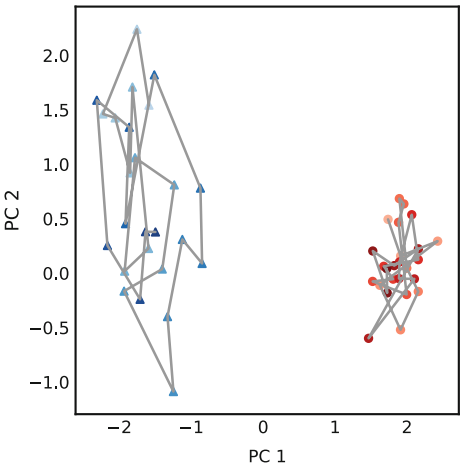

Fig. 3 An example of stepwise distance; the lines connect the first 25 consecutive chunks of two novels. Distances between points again represent semantic variance, but here the focus is on distances between consecutive chunks. Left: Wieringa, *Caesarion* (large distances). Right: Slee, *Fatale Liefde* (Fatal Attraction; small distances)

$$\text{stepwisedist}(T) = 1/(|T| - 1) \sum_{i=0}^{i<|T|-1} ||t_i - t_{i+1}||^2$$

This measure can detect the difference between small, gradual topic changes on the on hand, and large, sudden changes on the other, with respect to the linear progression of the text. See Fig. 3 for an illustration.

4.3 Outlier score

Aside from intra-textual variance, we can also consider inter-textual variance. The distance of a text to other texts can be operationalized by defining an *outlier score*. A simple approach is to measure the distance to the nearest neighboring novel (Ramaswamy et al. 2000):

$$\text{outlier}(T) = \min_{T' \in \text{ corpus}}^{T \neq T'} ||\mu_T - \mu_{T'}||$$

Each novel is represented as the centroid of its document vectors. See Fig. 4 for an illustration.

The corpus contains novels that are part of series, which may prevent them from being recognized as outliers with respect to the rest of the novels due to their similarity with each other. To correct for this, the outlier score could ignore the k nearest neighbors, with k being the number of novels in the longest series. However, in our experiments this did not improve the results, so we did not pursue this further.

4.4 Overlap score

A related idea is to measure the *overlap* of a novel's document vectors with those of other novels. We operationalize this by querying for the k-nearest neighbors around

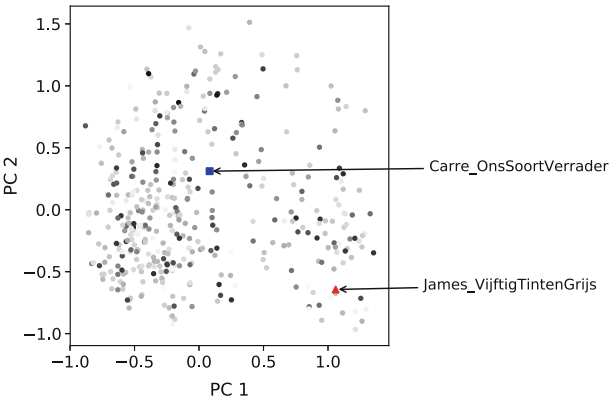

Fig. 4 An example of the outlier score. The novels are plotted in shades of gray corresponding to the literary rating (darker is higher rating). Two outliers are shown: James, *Fifty Shades of Grey* (Red triangle in lower right corner, overlapping with several other novels). Carre, *Our Kind of Traitor* (Blue square in the middle, far from other novels)

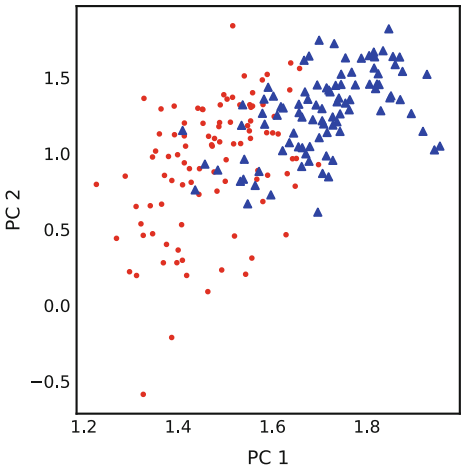

Fig. 5 An example of novels with overlap. Red dots: Royen, *Mannentester* (Man Tester); Blue triangles: Moelands, *Weerloos* (Defenseless). For an example of novels with no overlap, see Fig. 2

a novel's centroid, with k being the number of chunks in a novel, and returning the fraction of those neighbors which are part of other novels:

$$\text{overlap}(T) = 1/|T| \cdot |\{n_i^{\mu_T} : 0 < i < |T|\} \setminus T|$$

where n_i^T is the ith nearest neighboring chunk of μ_T. The neighbors can be computed efficiently with k-nearest neighbor algorithms (specifically, scikit-learn's BallTree). For an illustration, compare Figs. 2 and 5.

4.5 Evaluation

All of the above measures yield a single score for each novel, which can be correlated against their rating in the survey. See Table 3 for the results. For each of the variables, we find a considerable, statistically significant ($p < 0.05$) correlation in the expected direction.

Table 4 lists the top 5 novels for each variable. Note that some of the texts are short story collections, for which more intra-textual variance is to be expected.

Some titles can be found in several of the four complexity categories, at the same end of the spectrum. Carry Slee's two novels for young adults *Fatale Liefde* (Fatal Attraction) and *Bangkok Boy* both have a relatively low literary rating. Compared with the other novels in our corpus, both have relatively low variance and the stepwise distance between consecutive chunks of the text is also the smallest.

At the other end, Lanoye's novel *Sprakeloos* (Speechless), with a very high literary rating, has a very high intra-textual variance and also a relatively large stepwise distance. However, in terms of outlier score, this novel is average (1.877); in terms of overlap, it has a relatively high score (0.716), having the most overlap with a literary novel by Erwin Mortier and several literary novels by Dutch authors. Moreover, from the five titles with highest overlap, two have a high literary rating > 5, This goes against the intuition that the more literary a novel is, the less overlap it should have with other fiction, or that literary novels do not share similarities.

In the outlier category, at the bottom of the list, we find novels that are part of series, such as the three *Fifty Shades* novels by E. L. James. However, at the other end, with the highest score as outlier, we see novels that do not have a high score on literariness, by Dan Brown and J. K. Rowling, an exception to the tendency that the more a novel is an outlier, the more literary it is.

In conclusion, we find that our operationalizations of measuring structural and semantic complexity of novels provide some support for the hypothesis that literariness is characterized by a deviation from the norm. However, the results are far from a perfect correlation and there are exceptions, indicating that a reduction of literariness to semantic deviation is not warranted.

Table 3 Correlations of semantic complexity measures with literary ratings

Variable	Correlation (r)
Intra-textual variance	0.341*
Stepwise distance	0.431*
Outlier score	0.338*
Overlap score	−0.200*

*Indicates a statistically significant result with $p < 0.05$

Table 4 The top and bottom 5 novels for the complexity measures, with their literary rating

Label	Rating	Variance	Label	Rating	Outlier
Slee_FataleLiefde	4.141	8.219	James_VijftigTintenVrij	2.637	0.521
Rendell_Dief	4.212	8.516	James_VijftigTintenDonkerder	2.599	0.521
Slee_BangkokBoy	3.524	8.528	James_VijftigTintenGrijs	2.116	0.635
Groningen_Misleid	3.260	8.715	Collins_Vlammen	3.605	0.773
Voskuil_Buurman	6.053	8.780	Collins_Hongerspelen	3.460	0.773
...			...		
Buwalda_BonitaAvenue	5.844	15.683	Brown_VerlorenSymbool	3.646	2.882
Mak_ReizenZonderJohn	5.059	15.727	Jonasson_100-jarigeManDie	4.813	2.891
Dorrestein_Leesclub	4.977	15.820	Rowling_HarryPotterEn	3.826	2.905
Brokken_BaltischeZielen	5.579	16.393	Auel_LiedVanGrotten	3.659	2.924
Lanoye_Sprakeloos	6.373	16.415	Zwagerman_Duel	5.496	3.019
Label	Rating	Stepwisedist	Label	Rating	Overlap
Slee_FataleLiefde	4.141	13.880	Rosenboom_Mechanica	6.164	0.000
Slee_BangkokBoy	3.524	14.725	Campert_DagboekVanPoes	5.331	0.000
Groningen_Misleid	3.260	14.891	*King_EenmaligeZonde	4.010	0.000
Donoghue_Kamer	5.449	15.481	Grunberg_SelmonoskysDroom	6.125	0.000
Voskuil_Buurman	6.053	15.760	Meer_ZingenWaterPeen	5.017	0.000
...			...		
Dorrestein_Leesclub	4.977	27.692	Krauss_GroteHuis	5.990	0.759
Lanoye_HeldereHemel	5.826	28.279	Grisham_Wettelozen	3.914	0.784
Wieringa_PortretVanHeer	6.038	28.297	Meer_VrouwMetSleutel	5.349	0.800
Lanoye_Sprakeloos	6.373	30.014	Royen_Mannentester	3.180	0.847
Kooten_Verrekijker	4.962	30.208	*Sedaris_VanJeFamilie	4.389	0.935

Novels are labeled as 'Author_AbbreviatedTitle'. Texts marked with * are short story collections

5 Supervised predictive models

We apply linear models to to the task of predicting the rating for each document. Linear Support Vector Machines (SVM) are arguably the most popular predictive model for text classification; we use two closely related variants. For the BoW-models we use scikit-learn's (Pedregosa et al. 2011) SGDRegressor, which can be seen as an online version of SVM; i.e., trained incrementally instead of in a single batch. This is useful when both the number of samples and features is relatively large (in this case, the features consist of the whole vocabulary).

For the other models we apply an L_2-regularized Ridge model, which can be seen as a simpler version of linear SVM that has the same regularization but does not select support vectors from the training set. Support vectors are a subset of data points selected as representative to optimize the weights of the model.

We use 5-fold cross-validation for evaluating the predictive models, with the restriction that for each author, all of the chunks are in the same fold. This avoids the confounding factor of author-style being learned. The hyperparameters are tuned with crossvalidation on each training fold.

We report two evaluation metrics. R^2 (a percentage where the perfect score is 100) expresses the amount of variation in the original ratings that is explained by the model; this score is normalized. Root Mean Square Error (RMSE) gives the expected error for a prediction in the original scale of the ratings (1–7); this score is not normalized. A perfect result would have an error of 0. For example, when the RMS error for a model is 0.5, predicting the rating for a new novel should give a prediction that is on average 0.5 too high or too low with respect to its true rating.

Table 5 shows the main results with different document representations for predicting literary ratings from novels the respondents had read. We see that document embeddings outperform BoW-models, especially when both document embeddings are combined into a single model by concatenating their feature vectors.

The best result of the combined model is 52.2 R^2. Compare this to the result of van Cranenburgh and Bod (2017), who trained a model on the same task but used 1000 sentences per novel. Their model trained on textual features (without metadata features) achieves a score of 61.2 R^2. This means that our model is able to reproduce a large part of the performance with less than a tenth of the data (our 1000 word chunks contain 75 sentences on average).

Table 6 shows additional results with different variables from the survey. The literary ratings are substantially better predicted than the quality predictions. For the literary ratings there is a large difference between the read and not-read ratings, while for the quality ratings, the difference is much smaller for the R^2 score (the difference in RMSE is large, but this is explained by difference in range of the quality ratings, see below and Fig. 6). This suggest that the quality ratings are inherently difficult to predict from textual features.

Table 5 Scores for predicting the literary rating with each document representation versus combined

Model	R^2	RMSE
Bag of words, unigrams	35.5	0.786
Bag of words, bigrams	33.8	0.797
Topic model: LDA, unigrams	47.1	0.712
Paragraph Vectors (DBoW)	42.9	0.740
LDA and DBoW concatenated	**52.2**	**0.677**

Table 6 Predicting literariness versus quality, read and not read; using the combined LDA + DBoW model

Task	R^2	RMSE
Literary rating, read	52.2	0.677
Literary rating, not read	37.0	1.029
Quality rating, read	23.9	0.378
Quality rating, not read	21.6	0.919

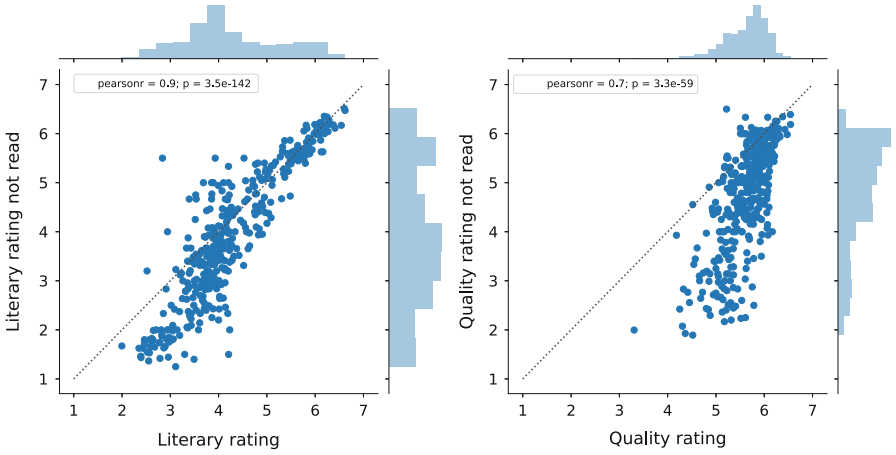

Fig. 6 The correlation of judgments on books the respondents had read and not read. The diagonal line represents a perfect correlation of ratings by readers and non-readers. Each dot represents a novel; dots above the line indicate novels rated higher by readers than non-readers, and vice versa for dots below the line

Note that for the judgments for books that respondents had not read, the number of ratings was not always sufficient for a reliable mean. Therefore part of the difference between the read and not-read predictions may be due to non-representative ratings. However, note that for both literary and quality ratings, the correlation between the 'read' and 'not read' ratings is both large and significant ($r = 0.9$ and $r = 0.7$, respectively, see Fig. 6). This implies that respondents do have certain expectations and opinions about books they have not yet read or do not intend to read and that these expectations are not completely opposite to actual readers' opinions. However, the 'read' literary ratings are substantially better predicted from the textual features. This supports the intuitive notion that readers rely on aspects of the text in making judgments on books they have read, while the judgments for books they have not read can only be influenced by the text indirectly and to a much lesser extent. The latter can occur with an author or novel famous for a particular writing style; e.g., the writing style of The Da Vinci Code and Fifty Shades of Grey has received wide attention, and it stands to reason that this has reached non-readers as well.

For the quality ratings, there is a marked difference between the range of ratings for the read and not-read novels. The quality ratings for read novels are compressed since almost all ratings are above 4 (on the border between bad and good). The lack of low ratings for quality is partly explained by the fact that the corpus consists exclusively of successful novels, but this leaves the question of the discrepancy between read and not-read quality ratings. Two kinds of biases may be at play here. On the one hand a selection and survivorship bias where readers pick or finish novels only when they are good enough; i.e., readers select novels they expect to be good, and readers may not finish novels that are not good enough. On the other hand

respondents may display choice-supportive bias by giving higher ratings for novels they have invested time in by finishing them.

It has been shown that the expectations about books not (yet) read are usually based on information from the media, publishers, reviewers, or fellow readers, or on where a book is placed in the book store and for instance the design of the cover of the book (Squires 2007; Verboord et al. 2015; Dixon et al. 2015). Our predictions of the ratings are solely based on the texts and do not include influences from sociological processes such as reviews and media hypes. A long-standing controversy about what informs the evaluation of literariness in relation to for example formation of the literary canon is described in a very accessible way by Fishelov (2008). He refers to the two opposite approaches as the beauty party versus the power party. The beauty party holds that intrinsic qualities of the text are responsible for a text being experienced as literary, whereas the power party are convinced that external social and cultural factors are the main factors responsible for the aesthetic values readers have. Publishers, reviewers, and so forth take decisions to label a certain text as literary, and these decisions are simply accepted and 'cloned' by readers without testing them on the text themselves. Fishelov advocates a combination of these two approaches, which he calls a 'dialogic approach' and which combines the ideas and tools of the different parties. Based on our research we not only see influences of the text but we also have strong indications that sociological processes play a role when readers rate books they read or did not read.

Figure 7 shows box plots of the prediction errors across the chunks of literary novels with the largest errors. It shows that for certain novels, the prediction is, on average, just right, while for other novels, there is a systematic bias towards over-, but mostly under-estimation. Furthermore, the range of predictions is wider for some novels. The novel at the bottom of the graph and thus the most underestimated is *The Sense of an Ending* by Julian Barnes. The Dutch translation, *Alsof het voorbij is*, was the novel in our corpus that received the highest mean score for literariness. Our survey ran from March to September 2013, and Barnes' novel won the prestigious Man Booker Prize in October 2011. We think it is very probable that this literary prize, which is very well-known and influential in the Netherlands, has affected the high ratings for this novel. This may imply that the fact that our model underestimates this novel based on the text could also be attributed to sociological influences at play in the actual reader judgments for this book.

Another underestimated novel is *De buurman* ('The neighbor', not available in English translation) by Dutch author J. J. Voskuil. Voskuil is renowned for his seven-volume novel *Het Bureau* ('The Institute', not available in English translation), which is currently identified as the third-longest novel in the world.[6] *Het Bureau* was enormously popular in the Netherlands starting in 1996, when the first volume was published, until long after volume seven appeared in 2000. The main topic of the novel is daily office life at a scholarly institute, and the writing style is misleadingly simple. One of its strong points is the dialogue. Contrary to what we usually see in literary novels, Voskuil included quite a lot of dialogue and he is especially admired for its realism and humor. *De buurman* was written in 2001

[6] See https://en.wikipedia.org/wiki/List_of_longest_novels.

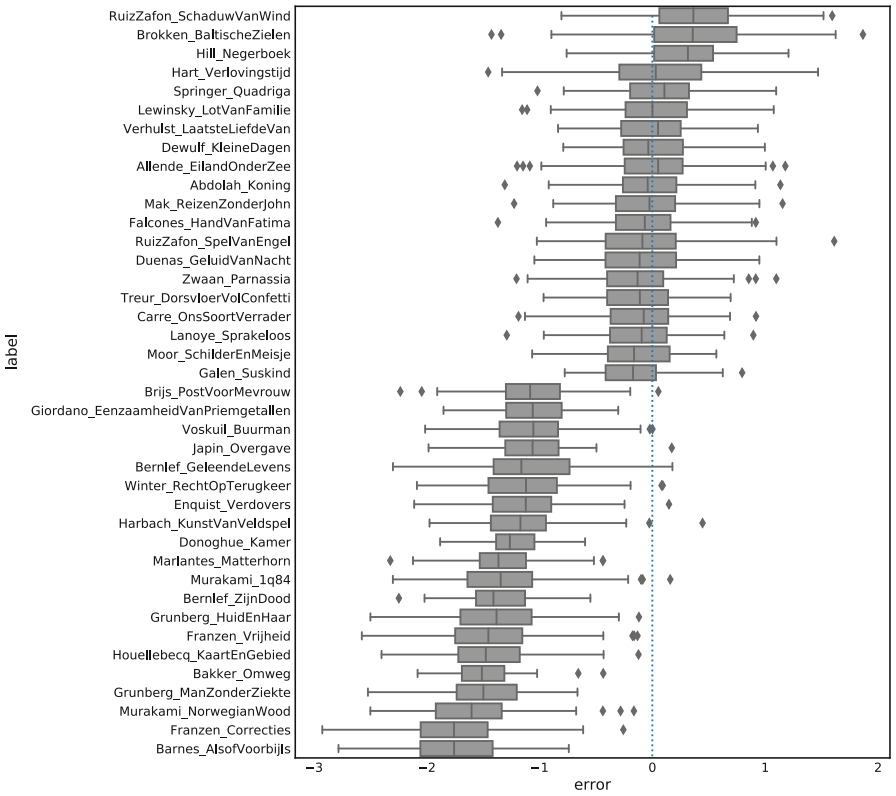

Fig. 7 Box plots of the prediction errors across chunks of selected literary novels. The boxes show the 1st quartile, median, and 3rd quartile (i.e., 50% of data points are within the box); the whiskers show the range of the values, except for outliers shown as dots. Novels are labeled as 'Author_AbbreviatedTitle'

and published posthumously in 2012 (Voskuil passed away in 2008), and uses the same writing style with perhaps even more dialogue than in his earlier novel. In the corpus we trained our model on, dialogue may be more prominent in other genres than the literary novel, and these genres are consistently rated as less literary. We hypothesize that Voskuil's reputation as a literary author informed the ratings for literariness, but that the exceptionally high amount of dialogue for a literary novel in *De buurman* may have led our model to partly underestimate it.

Our model has a tendency to underestimate novels. Only parts of some novels are overestimated. Figure 7 shows the twenty novels with the most overestimated parts. It is interesting that four of these are translations from Spanish. Our corpus has a total of seven novels translated from Spanish. This leads us to suggest that these Spanish novels represent a set of partly different stylistic literary conventions than the other novels in our corpus. The most interesting part is that this implies that these slightly different conventions are sufficiently recognizable in the Dutch translations of the Spanish originals to make them stand out here, which may be of interest from the perspective of Translation Studies.

6 Topic and keyword analysis

Section 4 showed how high-level intuitions of global novel structure can be tested with a geometric operationalization. However, the predictive models in the previous section show that the literariness of short text fragments is predictable as well. Subsequently we are interested in the specific, local textual features that the underlying computational models correlate with this general property of literariness. To get a more specific picture of the difference in language between novels we look at distinctive words associated with literariness. For the same reason we also look at words associated with prediction errors: if systematic prediction errors are correlated with similar words appearing in texts, those words may be associated with perceived higher or lower literariness, but not adequately captured by the model as such. In doing so, we are interested in words that are linked to content as well as in words that are linked to style. Broadly speaking, content words tend to be low to mid frequency terms, while function words tend to be high frequency terms.

Content words can be identified well in the topics of the LDA model, because each document is associated with a discrete list of weighted topics. Differences in these weights directly point to the importance of topics, which in turn are identified with a list of prominent words for the topic. We can inspect the topics associated with two subsets of the corpus by taking the mean topic weights for those subsets, and looking at the topics with the largest weights. However, since there may be overlap for topics that are associated with both subsets, we instead look at the topics which most strongly diverge between the two subsets. We report the topics including the names that were manually assigned, as reported in Jautze et al. (2016). Most topics identified a coherent theme, but some topics were found to be so specific for the novels of a particular author that the author name was assigned to the topic instead.

Function words are needed to give structure to sentences and carry little semantic meaning; this makes them useful in distinguishing writing style (cf. Burrows 1989). Distinctive function words can be identified by looking for words that stand out by having a higher frequency in a group of texts when contrasted with the frequencies in another group. This can be done with a statistical test such as the log-likelihood ratio as implemented in AntConc (Anthony 2005). We report keywords from the top 200 words with the highest keyness, after manually removing names.

6.1 The top 50 versus the bottom 50 by literary ratings

Data Novels rated least literary (range 2.12–3.21) versus novels rated most literary (range 5.77–6.62).

The most extreme contrast is between texts with the highest and lowest ratings. We find that several groups of function words are more frequent in literary novels: nouns and determiners, male pronouns, prepositions, and personal pronouns. Of these the nouns and determiners are associated with abstract concepts, while the pronouns refer to masculinity, and the personal pronouns indicate a more formal

Table 7 Distinctive topics across the novels with the top 50 and bottom 50 ratings

Diff.	Topics distinctive in bottom 50 novels	
− 0.070	t44: looks and parties	*vrouw glas jurk leuk uitzien* woman glass dress nice appear
− 0.067	t48: dialogues/colloquial language	*gewoon helemaal vertellen keer natuurlijk* normal completely tell time naturally
− 0.052	t31: (non-)verbal communication	*hand oog gezicht voelen aankijken* hand eye face feel look-at
− 0.048	t46: author: Kinsella/Wickham	*mam opeens gewoon krijgen voelen* mum suddenly normal receive feel
− 0.044	t23: settling down	*leven huis kind vrouw jaar* live house child woman year
Diff.	**Topics distinctive in top 50 novels**	
0.030	t42: time, life and death	*dag één slechts straat tijd* day one only street time
0.037	t26: nature/life	*licht oog voelen liggen leven* light eye feel lie live
0.039	t41: writers	*boek schrijven lezen één verhaal* book write read one story
0.039	t1: self-development	*leven tijd mens moment blijven* live time human moment remain
0.052	t29: music/performance/misc	*beginnen muziek spelen keer eerst* begin music play time first

The value on the left is the mean weight difference (top50 − bottom50)

somewhat distancing use of language (formal 'you', formal 'your', 'they', 'we', 'one', 'he', 'our', 'she').

Because the signal we find is rather weak we refrain from making strong claims about what sets literary vocabulary apart from lesser literary language. It is tempting maybe to interpret these observations as indicating that literary language is associated with more formal and disinterested description, and that the preference for abstract notions suggests an intellectual horizon, while the propensity to use personal pronouns is more indicative for an interest in the 'other' than for the 'self.' This would then contrast to the rather more concrete notions of lesser literary texts that focus primarily on the self of the protagonist and her self-reflexive immediate social relations as she is immersed in hedonistic social events. Although being in accordance perhaps with some intuitions, such characterizations should be taken as tentative conjecture, and as an interesting challenge for further investigation at best, given the relative weak signal. However, our approach does demonstrate that indeed there are avenues to interrogate machine learning models that are able to predict

Table 8 Distinctive topics across the most and least literary novels

Diff.	Topics distinctive in bottom 50 literary novels	
− 0.043	t31: (non-)verbal communication	*hand oog gezicht voelen aankijken* hand eye face feel look-at
− 0.031	t3: author: Auel	*grot paard vrouw mens man* cave horse woman human man
− 0.024	t23: settling down	*leven huis kind vrouw jaar* live house child woman year
− 0.018	t11: children	*kind moeder mama baby papa* child mother mama baby daddy
− 0.010	t44: looks and parties	*vrouw glas jurk leuk uitzien* woman glass dress nice appear
Diff.	**Topics distinctive in top 50 literary novels**	
0.012	t30: education	*school jongen meisje eerste leerling* school boy girl first student
0.012	t1: self-development	*leven tijd mens moment blijven* live time human moment remain
0.020	t42: time, life and death	*dag één slechts straat tijd* day one only street time
0.035	t41: writers	*boek schrijven lezen één verhaal* book write read one story
0.046	t29: music/performance/misc	*beginnen muziek spelen keer eerst* begin music play time first

The value on the left is the mean weight difference (top50lit − bottom50lit)

literariness judgments of a general audience of readers well, and that it is possible to query such models for what terms they associate with higher or lower perceived literariness.

It must be noted, however, that the make up of the corpus is such that the subsets do not differ exclusively in terms of their literary ratings. Other important variables are genre, author gender, and whether the novel is translated. This can be seen from the topics in Table 7, which clearly show the influence of chick lit novels in the bottom 50. This raises the question of whether we are uncovering intrinsic aspects of literariness, or just various incidental dataset biases (such as stylistic conventions associated with less literary genres). For an in depth investigation of biases concerning author gender, cf. Koolen (2018). To avoid this issue we now turn to a subset of only literary novels.

6.2 Literary versus highly literary texts

Data Only texts marketed as literary novels by the publisher. Lowest rated (range 3.05–4.89) versus highest rated (range 5.76–6.62).

To control for the factor of genre, we now compare the top 50 most literary novels to the literary novels with the lowest ratings. This excludes romantic novels and thrillers. The distinctive topics for these subsets are given in Table 8. The keywords distinctive for the top literary novels are as follows:

- Personal pronouns: *hij* (he), *zijn* (his), *wij* (we), *mij* (me), *men* (one), *zij* (she), *zelf* (self), *ge* (formal you, archaic/Flemish)
- Prepositions: *van* (of), *in* (in), *zonder* (without)
- Determiners: *de* (the), *een* (a)
- Conjunctions: *of* (or), *doch* (yet, archaic)
- Adverbs: *ook* (also), *nog* (still), *al* (already)
- Nouns typically referring to male persons: *kok* (cook), *schrijver* (author), *vizier* (vizier), *loper* (walker), *opperhoofd* (chief), *luitenant* (lieutenant), *magistraat* (magistrate), *houthakker* (lumber), *sjeeg* (sheikh), *cowboy*, *schilder* (painter), *kaptein* (captain), *kapitein* (captain), *kokkie* (cook)

And distinctive keywords for the literary novels with the lowest ratings can be summarized as follows:

- Personal pronouns: *ze* (she), *haar* (her), *mijn* (my), *me* (me), *hen* (they), *we* (we), *je* (you), *ik* (I)
- Prepositions: *naar* (to, direction), *om* (to, indicating time or introducing a relative clause)
- Conjunctions: *en* (and), *voordat* (before)
- Adverbs: *snel* (fast)
- Nouns referring to intimate and family relations: *baby*, *moeder* (mother), *god*, *papa* (dad), *kinderen* (children), *oma* (grandma), *vrouw* (wife), *mama* (mom), *mutti* (mommy), *zus* (sister), *zussen* (sisters), *famillie* (family), *mannen* (men), *mam* (mum).

What seemed at best conjecture when looking at the full data corpus turns out to become a stronger signal when we only look at literary novels. Higher rated literary novels involve more formal language and deal with the affairs and exploits of male characters in a wider society setting. Life, death, writing, and intellectual development are categories associated with these literary novels. This is sharply contrasted with the intimate and family-oriented events of female characters described in lower rated literature in more informal and self-oriented language focusing on looks, parties, and children.

We also find an interesting linguistic contrast concerning full and reduced variants of pronouns (see Table 9), which come up as keywords on both sides. Linguistically, the full pronouns are considered strong emphatic, while the reduced pronouns are weak unemphatic; the distinction relates to contrast and salience of

Table 9 Full and reduced personal pronouns in Dutch

	Full	Reduced
1st sg	ik, mij	me
2nd sg	jij, jou	je
3rd sg fem	zij, haar	ze
1st pl	wij, ons	we
3rd pl	zij, hen/hun	ze

The first column shows subject and object forms

discourse referents (Kaiser 2010). In some situations, one or the other is required, while in other situations, both are permitted, making the contrast a stylistic choice. Among other differences, the reduced pronouns are more informal and are required in fixed expressions such as *dank je* (thank you), whereas the full pronouns can be used for emphasis or refer to a less salient referent; the full version is required when expressing contrast. The stylistic aspects of the Dutch reduced pronouns warrant further study.

6.3 Over-versus underestimated chunks

Data Only novels with a high literary rating (rating > 5, 98 novels). Divided by prediction error with the DBoW model into underestimated (predicted $-$ actual < -0.5, 7855 chunks), overestimated (predicted $-$ actual > 0.5, 218 chunks), and small error (rest, 3832 chunks).

A similar procedure can be applied to the specific chunks of texts for which the model predictions exhibited a large or small error. This can serve to highlight which aspects of literariness the model has successfully learned, and which aspects cause it to make systematic errors. We divide the chunks into three categories: underestimated, overestimated, and the rest with a smaller error. We set the boundary of these categories at 0.5. That is, for a novel with a rating of 6, a prediction of 5.5 or lower would be underestimated, and 6.5 or higher would be filed in the category of overestimated chunks.

See Table 10 for the results with LDA topics. The underestimated chunks are associated with topics related to dialogue, physical attack, and communication—indicating that dialogue and violence have been associated with less literary novels. The overestimated chunks are associated with topics related to family, church, and life and death.

Once again we also used AntConc to identify keywords that stand out when texts in these categories are contrasted. To focus on the largest contrast we compare the under- and overestimated chunks (as opposed to including the middle category in comparisons). Figure 8 shows box plots of the frequencies of these keywords in each subset. Although again prudence is warranted, it does seem that these findings corroborate what we found in the previous tests. Female pronouns ('she' and 'her', possibly indicative of female lead characters or a majority of female characters) are more frequent in underestimated chunks, as are first and second person pronouns ('I', 'you', 'me', 'myself', 'we', 'your'). This again suggests that novels with a high 'female make up' and a strong focus on the self of the subject are evaluated as less

Table 10 The five most diverging topic weights for under- and overestimated chunks

Diff.	Topics distinctive in underestimated chunks	
− 0.030	t48 dialogues/colloquial language	*gewoon helemaal vertellen keer natuurlijk* normal completely tell occasion naturally
− 0.017	t25 physical attack	*hand hoofd arm man proberen* hand head arm man try
− 0.017	t31 (non-)verbal communication	*hand oog gezicht voelen aankijken* hand eye face feel look-at
− 0.016	t1 self-development	*leven tijd mens moment blijven* live time human moment stay
− 0.012	t37 military	*soldaat luitenant leger twee krijgen* soldier lieutenant army two receive
Diff.	**Topics distinctive in overestimated chunks**	
0.018	t43 jewishness/world war II	*jood joods mens duits twee* jew jewish human german two
0.019	t2 family	*vader moeder kind jaar zoon* father mother child year son
0.03	t35 church	*kerk man vrouw dominee priester* church man woman preacher priest
0.04	t42 time, life and death	*dag één slechts straat tijd* day one only street time
0.093	t36 international politics	*land jaar oorlog amerikaans stad* land year war american city

The value on the left is the weight difference (over − under)

literary. Also the presence of perception and cognition related words (‘know’, ‘think’, ‘look’, ‘feel’, ‘felt’, ‘meant’, ‘remember’, ‘realize’) that indicate an explicit pre-occupation with how the subject is experiencing events may be indicative of this pre-occupation with the self in novels that are judged less literary. In underestimated chunks we also find a significantly higher use of verbs related to dialogue (‘say’, ‘says’, ‘said’, ‘asked’) and of negations (*niet* ‘not’, *geen* ‘no’ as determiner). Conversely, and again corroborating what has been found in the previous tests, we see that male personal pronouns (‘he’, ‘him’) and nouns referring to male positions in society (‘coach’, ‘shah’, ‘mister’, ‘batsman’, ‘captain’) are associated with chunks being overestimated, thus signaling a text’s pre-occupation with the social status of men as a potential marker for literariness.

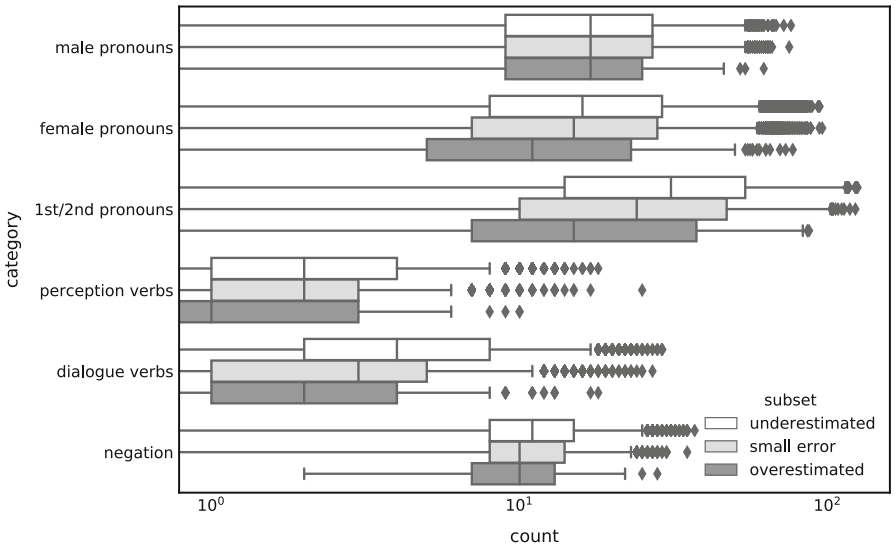

Fig. 8 Box plots of the mean keyword frequencies in under- and overestimated chunks. The x-axis has a logarithmic scale

7 Discussion and conclusion

We have shown to what degree the perceived literariness of novels is reflected in their texts with neural embeddings and topic models. Compared to previous work we have shown that literariness can be predicted well even when the model is presented with a much smaller quantity of text of 2–3 pages. This provides a precise lower bound on how informative the text of a novel is when the task is to predict its perceived literariness, without using any text-external (e.g., sociological) knowledge. The result is precise in the sense that to the extent the predicted ratings are correct, they are objectively and reproducibly so; the result is a lower bound because it is likely that a more sophisticated model could perform even better.

Aside from quantifying the predictability of literariness, we have attempted to explore factors that may determine literariness by exploring various uses of distance measures. This allowed us to show how various forms of semantic complexity are associated with literariness, providing some support for the hypothesis that literariness is a semantic deviation from the norm, although it is clear that literariness cannot be reduced to such deviation.

Topic and keyword analysis suggests several distinctive themes and stylistic differences across novels rated from highly literary to not at all literary. However, not all differences seem to represent intrinsic aspects of literariness. Several themes and keywords are either associated with literariness through genre or bias the predictions in the wrong direction. High frequency keywords related to gender, dialogue, narrative perspective, and negation are associated with different levels of literariness. Some of these clearly represent incidental biases (e.g., the

overrepresentation of male protagonists in literary novels), while others could well reflect intrinsic markers of literariness (e.g., a more distanced narrative perspective)

The paragraph vector model is good at creating a systematic vector space of documents and words, and these distances can be meaningfully queried, as demonstrated by the semantic complexity measures in Sect. 4. However, inspecting the influence of specific language is harder, because the influence of stylistic differences on the vectors is rather opaque, and its dense representations are not interpretable in terms of a discrete list of topics. The role of topic and style in paragraph vectors is an interesting subject for future research. Topic models are still better at identifying discrete, coherent themes, and can play a complementary role.

As noted in Sect. 2, the survey data consists of literary ratings per novel, while we set out to investigate intra-textual variance in literariness by looking at chunks of novels. Confirming a direct relationship between textual features and literariness would require a more detailed survey in which participants provide ratings with a finer granularity on this level of novel chunks. However, our findings already suggest that such associations between textual features and literariness obtain, and point to particular stylistic aspects that merit further research.

Another limitation is the choice to predict average literary ratings. A more sophisticated model could model preferences of individual readers, and could take into account the fact that each novel has been read and rated by a particular subset of the respondents.

Our results imply that there are structural factors associated with literariness (as seen in the semantic complexity measures), and confirm social prejudices about genre. However, the main implication is that the greater part of factors affecting judgments of literariness are explicable in (distributed) bag-of-words terms, even *within* the literary genre and among novels with higher literary ratings.

Our work has uncovered further details on perceptions of literariness. However, it is clear that more methods and data are needed to fully understand the range of stylistic devices that literary language employs to distinguish itself in the perception of readers. Specifically, the bag-of-words information employed by the models cannot be said to be equivalent to writing style, and disentangling the role of aspects such as style, topics, plot structure, narrative pace, etc., remains an open challenge for future work.

Acknowledgements We are grateful to Corina Koolen and the anonymous reviewers for valuable comments on this paper.

References

Anthony, L. (2005). Antconc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom. In *Proceedings of the 2005 IEEE international professional communication conference* (pp. 729–737). IEEE.

Ashok, V., Feng, S., & Choi, Y. (2013). Success with style: Using writing style to predict the success of novels. In *Proceedings of EMNLP* (pp. 1753–1764). http://aclweb.org/anthology/D13-1181.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bouma, G., Van Noord, G., & Malouf, R. (2001). Alpino: Wide-coverage computational analysis of Dutch. *Language and Computers*, *37*(1), 45–59.

Brennan, T. (2017). The digital-humanities bust. In *The Chronicle of Higher Education*, October 20. http://www.chronicle.com/article/The-Digital-Humanities-Bust/241424. Accessed 28 Oct 2017.

Burrows, J. F. (1989). 'An ocean where each kind...': Statistical analysis and some major determinants of literary style. *Computers and the Humanities*, *23*(4–5), 309–321.

Clement, T., Steger, S., Unsworth, J., & Uszkalo, K. (2008). How not to read a million books. In *Personal page at institutional site*. University of Virgina, October 2008. http://people.virginia.edu/~jmu2m/hownot2read.html. Accessed 28 Oct 2017.

Crosbie, T., French, T., & Conrad, M. (2013). Towards a model for replicating aesthetic literary appreciation. In *Proceedings of the 5th workshop on semantic web information management* (p. 8). https://doi.org/10.1145/2484712.2484720

Dai, A. M., Olah, C., & Le, Q. V. (2015). Document embedding with paragraph vectors. arXiv e-print arxiv:1507.07998.

Dixon, P., Bortolussi, M., & Mullins, B. (2015). Judging a book by its cover. *Scientific Study of Literature*, *5*(1), 23–48. https://doi.org/10.1075/ssol.5.1.02dix.

Fish, S. (2012). Mind your p's and b's: The digital humanities and interpretation. *The New York Times*, January 23. http://opinionator.blogs.nytimes.com/2012/01/23/mind-your-ps-and-bs-the-digital-humanities-and-interpretation/. Accessed 28 Oct 2017.

Fishelov, D. (2008). Dialogues with/and great books: With some serious reflections on Robinson Crusoe. *New Literary History*, *39*(2), 335–353.

Herrmann, J. B., van Dalen-Oskam, K., & Schöch, C. (2015). Revisiting style, a key concept in literary studies. *Journal of Literary Theory*, *9*(1), 25–52.

Jautze, K., van Cranenburgh, A., & Koolen, C. (2016). Topic modeling literary quality. In *Digital humanities 2016: Conference abstracts* (pp. 233–237), Krákow, Poland. http://dh2016.adho.org/abstracts/95.

Kaiser, E. (2010). Effects of contrast on referential form: Investigating the distinction between strong and weak pronouns. *Discourse Processes*, *47*(6), 480–509.

Kestemont, M., & Stutzmann, D. (2017). Script identification in medieval latin manuscripts using convolutional neural networks. In *Digital Humanities 2017 Book of Abstracts, ADHO, Montreal* (pp. 283–285). https://dh2017.adho.org/abstracts/078/078.pdf.

Kirschenbaum, M. (2014). What is digital humanities and what's it doing in english departments? *Differences*, *25*(1), 46–63. https://doi.org/10.1215/10407391-2419997.

Koolen, C. (2018). Reading beyond the female: The relationship between perception of author gender and literary quality. *Ph.D. thesis*, University of Amsterdam. http://hdl.handle.net/11245.1/cb936704-8215-4f47-9013-0d43d37f1ce7.

Lau, J. H., & Baldwin, T. (2016). An empirical evaluation of doc2vec with practical insights into document embedding generation. In *Proceedings of the representation learning for NLP workshop* (pp 78–86). http://aclweb.org/anthology/W16-1609.

Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of ICML* (pp. 1188–1196). http://jmlr.org/proceedings/papers/v32/le14.pdf.

Lei, T., Barzilay, R., & Jaakkola, T. (2016). Rationalizing neural predictions. In *Proceedings of EMNLP* (pp. 107–117). http://aclweb.org/anthology/D16-1011.

Louwerse, M. (2004). Semantic variation in idiolect and sociolect: Corpus linguistic evidence from literary texts. *Computers and the Humanities*, *38*(2), 207–221.

Louwerse, M., Benesh, N., & Zhang, B. (2008). Computationally discriminating literary from non-literary texts. In S. Zyngier, M. Bortolussi, A. Chesnokova, & J. Auracher (Eds.), *Directions in empirical literary studies: In honor of Willie Van Peer* (pp. 175–191). Amsterdam: John Benjamins Publishing Company.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of ACL-HLT* (pp. 142–150). http://aclweb.org/anthology/P11-1015.

Maharjan, S., Arevalo, J., Montes, M., González, F. A., & Solorio, T. (2017). A multi-task approach to predict likability of books. In *Proceedings of EACL* (pp. 1217–1227). http://aclweb.org/anthology/E17-1114.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu. Accessed 15 Oct 2017.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Ramaswamy, S., Rastogi, R., & Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. *ACM Sigmod Record*, *29*, 427–438. https://doi.org/10.1145/335191.335437.

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks* (pp. 45–50). http://is.muni.cz/publication/884893/en.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trustyou?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778.

Sculley, D., & Pasanek, B. M. (2008). Meaning and mining: The impact of implicit assumptions in data mining for the humanities. *Literary and Linguistic Computing*, *23*(4), 409–424. https://doi.org/10.1093/llc/fqn019.

Squires, C. (2007). *Marketing literature: The making of contemporary writing in Britain*. Basingstoke: Palgrave Macmillan. 10.1057/9780230593008.

Underwood, T. (2013). We don't already understand the broad outlines of literary history. In *The stone and the shell (academic blog)*, February 8. https://tedunderwood.com/2013/02/08/we-dont-already-know-the-broad-outlines-of-literary-history/. Accessed 15 Oct 2017.

Underwood, T. (2015). The literary uses of high-dimensional space. *Big Data & Society 2*(2). http://bds.sagepub.com/content/2/2/2053951715602494.

van Cranenburgh, A., & Bod, R. (2017). A data-oriented model of literary language. In *Proceedings of EACL* (pp. 1228–1238). http://aclweb.org/anthology/E17-1115.

van Cranenburgh, A., & Koolen, C. (2015). Identifying literary texts with bigrams. In *Proceedings of workshop computational linguistics for literature* (pp. 58–67). http://aclweb.org/anthology/W15-0707.

Verboord, M., Kuipers, G., & Janssen, S. (2015). Institutional recognition in the transnational literary field, 1955–2005. *Cultural Sociology*, *9*(3), 447–465. https://doi.org/10.1177/1749975515576939.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. J., & Hovy, E. H. (2016). Hierarchical attention networks for document classification. In *Proceedings of HLT-NAACL* (pp. 1480–1489). http://aclweb.org/anthology/N16-1174.