university of groningen

University Medical Center Groningen

University of Groningen

Identifying literary texts with bigrams

van Cranenburgh, Andreas; Koolen, Corina

Published in:
Proceedings of CLFL

DOI:
10.3115/v1/W15-0707

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2015

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
van Cranenburgh, A., & Koolen, C. (2015). Identifying literary texts with bigrams. In *Proceedings of CLFL* (pp. 58-67). Association for Computational Linguistics (ACL). https://doi.org/10.3115/v1/W15-0707

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 26-12-2020

Identifying Literary Texts with Bigrams

Andreas van Cranenburgh[*†]
[*]Huygens ING
Royal Dutch Academy of Sciences
The Hague, The Netherlands
Andreas.van.Cranenburgh@huygens.knaw.nl

Corina Koolen[†]
[†]Institute for Logic, Language and Computation
University of Amsterdam
The Netherlands
C.W.Koolen@uva.nl

Abstract

We study perceptions of literariness in a set of contemporary Dutch novels. Experiments with machine learning models show that it is possible to automatically distinguish novels that are seen as highly literary from those that are seen as less literary, using surprisingly simple textual features. The most discriminating features of our classification model indicate that genre might be a confounding factor, but a regression model shows that we can also explain variation between highly literary novels from less literary ones within genre.

1 Introduction

> The prose, plot, [the] characters, the sequence of the events, the thoughts that run in Tony Websters mind, big revelation in the end . . . They are all part of the big beautiful ensemble that delivers an exceptionally nice written novella. — (from a review on Goodreads of Julian Barnes, *A Sense of an Ending*)

However much debated the topic of literary quality is, one thing we do know: we cannot readily pinpoint what 'literary' means. Literary theory has insisted for a number of years that it lies mostly outside of the text itself (cf. Bourdieu, 1996), but this claim is at odds with the intuitions of readers, of which the quote above is a case in point. Publishers, critics, and literary theorists all influence the opinions of readers, but nevertheless, in explaining the sense of rapture or awe they experience, they will choose textual elements to refer to. In our project,[1] we try to find whether novels that are seen as literary have certain textual characteristics in common, and if so, what meaning we can assign to such commonalities. In other words, we try to answer the following question: are there particular textual conventions in literary novels that contribute to readers judging them to be literary?

In this paper, we show that there are indeed textual characteristics that contribute to perceived literariness. We use data from a large survey conducted in the Netherlands in 2013, in which readers were asked to rate novels that they had read on a scale of literariness and of general quality (cf. section 2). We show that using only simple bigram features (cf. section 3), models based on Support Vector Machines can successfully separate novels that are seen as highly literary from less literary ones (cf. section 4). This works with both content and style related features of the text. Interestingly, general quality proves harder to predict. Interpretation of features shows that genre plays a role in literariness (cf. section 5), but results from regression models indicate that the textual features also explain differences within genres.

2 Survey Data and Novels

During the summer of 2013, the Dutch reading public was asked to give their opinion on 401 novels published between 2007 and 2012 that were most often sold or borrowed between 2009 and 2012. This list was chosen to gather as many ratings as possible

[1]The Riddle of Literary Quality,
cf. http://literaryquality.huygens.knaw.nl

Proceedings of NAACL-HLT Fourth Workshop on Computational Linguistics for Literature, pages 58–67,
Denver, Colorado, June 4, 2015. ©2015 Association for Computational Linguistics

	Original	Translated
Thrillers	0	31
Literary thrillers	26	29
Literary fiction	27	33

Table 1: The number of books in each category. These categories were assigned by the publishers.

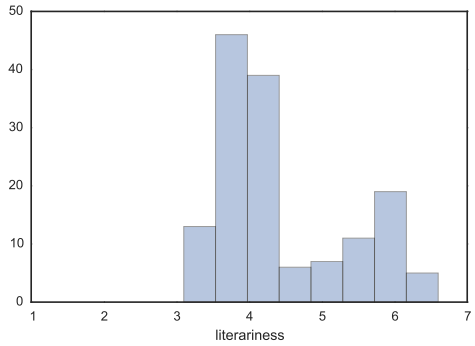

Figure 1: A histogram of the mean literary ratings.

(less popular novels might receive too few ratings for empirical analysis), and to ensure that readers were not influenced too much by common knowledge on their canonisation (this is less likely for more recent books). About 13,000 people participated in the survey. Participation was open to anyone. Participants were asked, among other things, to select novels that they had read and to rate them on two scales from 1–7: literariness (not very literary–very literary) and general quality (bad–good). These two were distinguished because a book that is not literary can still be considered to be a good book, because it is suspenseful or funny for instance; conversely, a novel that is seen as literary can still be considered to be bad (for instance if a reader does not find it engaging), although we found no examples of this in our results. No definition was given for either of the two dimensions, in order not to influence the intuitive judgments of participants. The notion of literariness in this work is therefore a pretheoretical one, directly reflecting the perceptions of the participants. In this work we use the mean of the ratings of each book.

The dataset used in this paper contains a selection of 146 books from the 401 included in the survey; see Table 1 and 2. Both translated and original (Dutch) novels are included. It contains three genres, as indicated by the publisher: literary novels, literary thrillers and thrillers. There are no Dutch thrillers in the corpus. Note that these labels are ones that the publishers have assigned to the novels. We will not be using these labels in our experiments—save for one where we interpret genre differences—we base ourselves on reader judgements. In other words: when we talk about highly literary texts, they (in theory) could be part of any of these genres, as long as readers judged them to be highly literary.

3 Experimental setup

Three aspects of a machine learning model can be distinguished: the target of its predictions, the data which predictions are based on, and the kind of model and predictions it produces.

3.1 Machine Learning Tasks

We consider two tasks:

1. Literariness
2. Bad/good (general quality)

The target of the classification model is a binary classification whether a book is within the 25 % judged to be the most literary, or good. Figure 1 shows a histogram of the literary judgments. This cutoff divides the two peaks in the histogram, while ensuring that the number of literary novels is not too small.

A more difficult task is to try to predict the average rating for literariness of each book. This not only involves the large differences between thrillers and literary novels, but also smaller differences within these genres.

3.2 Textual Features

The features used to train the classifier are based on a bag-of-words model with relative frequencies. Instead of single words we use word bigrams. Bigrams are occurrences of two consecutive words observed in the texts. The bigrams are restricted to those that occur in between 60 % and 90 % of texts used in the model, to avoid the sparsity of rare bigrams on the one hand, and the most frequent function bigrams on

	Original	Translated
literature	Bernlef, Dis, Dorrestein, Durlacher, Enquist, Galen, Giphart, Hart, Heijden, Japin, Kluun, Koch, Kroonenberg, Launspach, Moor, Mortier, Rosenboom, Scholten, Siebelink, Verhulst, Winter.	Auel, Avallone, Baldacci, Binet, Blum, Cronin, Donoghue, Evans, Fragoso, George, Gilbert, Giordano, Harbach, Hill, Hodgkinson, Hosseini, Irving, James, Krauss, Lewinsky, Mastras, McCoy, Pick, Picoult, Rosnay, Ruiz Zafn, Sansom, Yalom.
literary thrillers	Appel, Dijkzeul, Janssen, Noort, Pauw, Terlouw, Verhoef, Vermeer, Visser, Vlugt	Coben, Forbes, French, Gudenkauf, Hannah, Haynes, Kepler, Koryta, Lackberg, Larsson, Lckberg, Nesbo, Patterson, Robotham, Rosenfeldt, Slaughter, Stevens, Trussoni, Watson.
thrillers		Baldacci, Clancy, Cussler, Forsyth, Gerritsen, Hannah, Hoag, Lapidus, McFadyen, McNab, Patterson, Roberts, Rose.

Table 2: Authors in the dataset

the other. No limit is placed on the total number of bigram features. We consider two feature sets:

content bigrams: Content words contribute meaning to a sentence and are thus topic related; they consist of nouns, verbs, adjectives, and adverbs. Content bigrams are extracted from the original tokenized text, without further preprocessing.

style bigrams: Style bigrams consist of function words, punctuation, and part-of-speech tags of content words (similar to Bergsma et al. 2012). In contrast with content words, function words determine the structure of sentences (determiners, conjunctions, prepositions) or express relationships (pronouns, demonstratives). Function words are identified by a selection of part-of-speech tags and a stop word list. Function words are represented with lemmas, e.g., auxiliary verbs appear in uninflected form. Lemmas and part-of-speech tags were automatically assigned by the Alpino parser.[2]

3.3 Models

All machine learning experiments are performed with `scikit-learn` (Pedregosa et al., 2011). The classifier is a linear Support Vector Machine (SVM) with regularization tuned on the training set. The cross-validation is 10-fold and stratified (each fold has a distribution of the target class that is similar to that of the whole data set).

For regression the same setup of texts and features is used as for the classification experiments, but the machine learning model is a linear Support Vector Regression model.

4 Results

Before we train machine learning models, we consider a dimensionality reduction of the data. Figure 2 shows a non-negative matrix factorization of the style bigrams. In other words, this is a visualization of a decomposition of the bigram counts, without taking into account whether novels are literary or not (i.e., an unsupervised model). Notice that most of the non-literary novels (red) cluster together in one corner, while the literary books (blue) show more variation. When content bigrams are used, a similar cluster of non-literary books emerges, but interestingly, this cluster only consists of translated works. With style bigrams this does not occur.

This result seems to suggest that non-literary books are easier to recognize than literary books, since the literary novels show more variation. However, note that this decomposition present just one way to summarize and visualize the data. The classification

[2]Cf. `http://www.let.rug.nl/vannoord/alp/Alpino/`

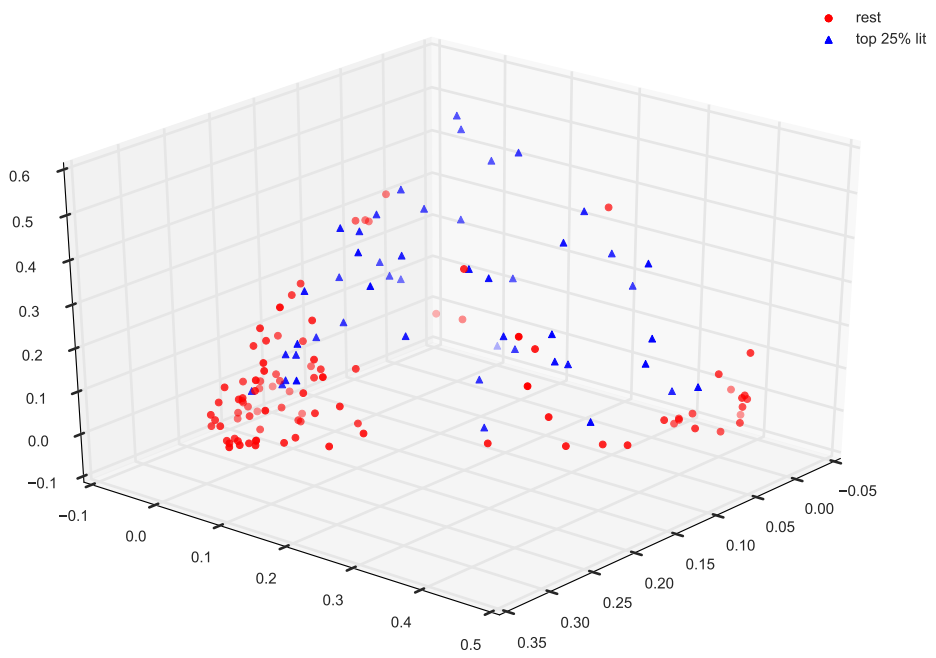

Figure 2: Non-negative matrix factorization based on style bigrams (literary novels are the blue triangles).

Features	Literary	Bad/good
Content bigrams	90.4	63.7
Style bigrams	89.0	63.0

Table 3: Classification accuracy (percentage correct).

Features	Literary	Bad/Good
Content bigrams	61.3 (0.65)	33.5 (0.49)
Style bigrams	57.0 (0.67)	22.2 (0.52)

Table 4: Evaluation of the regression models; R^2 scores (percentage of variation explained), root mean squared error in parentheses (1–7).

model, when trained specifically to recognize literary and non-literary texts, can still identify particular discriminating features.

4.1 Classification

Table 3 shows the evaluation of the classification models. The content bigrams perform better than the style bigrams. The top-ranked bigram features of the model for literary classification are shown in Table 5.

If we look only at the top 20 bigrams that are most predictive of literary texts according to our model and plot how often they occur in each genre as specified by the publishers, we see that these bigrams occur significantly more often in literary texts; cf. the plot in Figure 4. This indicates that there are features specific to literary texts, despite the variance among literary texts shown in Figure 2.

When trained on the bad/good dimension, the classification accuracy is around 60 %, compared to around 90 % for literariness, regardless of whether the features are about content or style bigrams. This means that the bad/good judgments are more difficult to predict from these textual features. This is not due to the variance in the survey responses themselves. If literariness were a more clearly defined concept for the survey participants than general quality, we would expect there to be less consensus and thus more variance on the latter dimension. But this is not what we find; in fact the mean of the standard deviations of the bad/good responses is lower than for the literariness responses (1.08 vs. 1.33). Rather, it is likely that the bad/good dimension depends on higher-level, plot-related characteristics, or text-extrinsic social factors.

4.2 Regression

The regression results cannot be evaluated with a simple ‘percentage correct’ accuracy metric, because

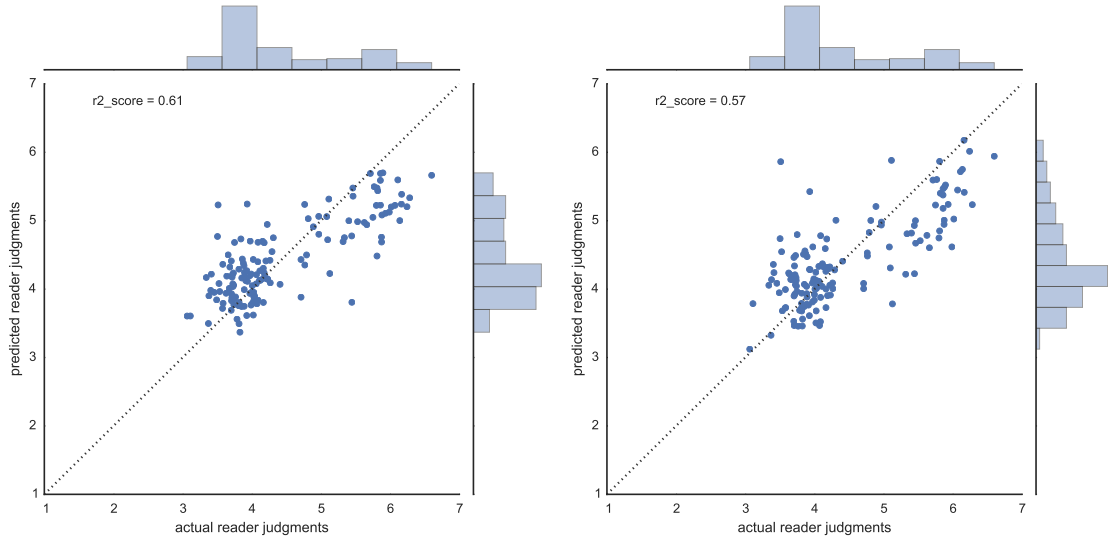

Figure 3: Regression results for predicting literary judgments with content bigrams (left) and style bigrams (right).

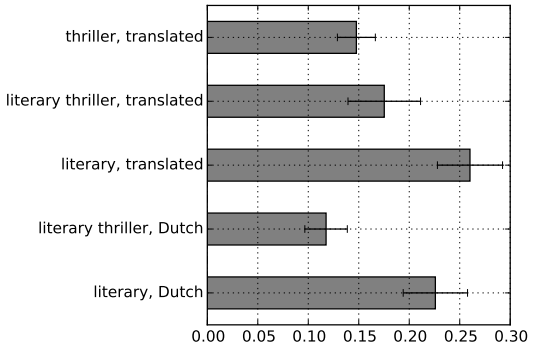

Figure 4: A barplot of the number of occurrences of the top 20 most important literary features (cf. Table 5) across the genres given by the publisher (error bars show 95 % confidence interval).

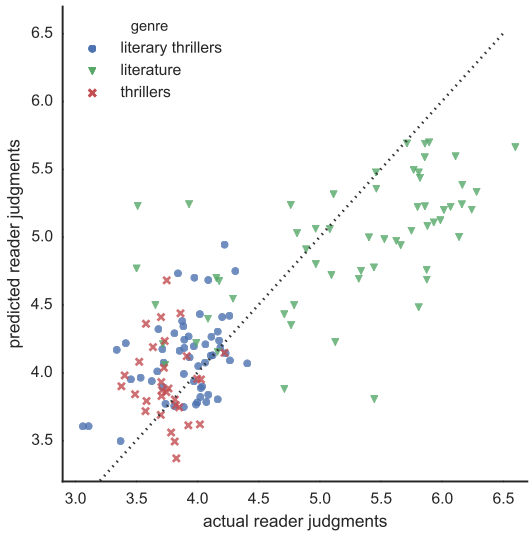

Figure 5: Regression results for predicting literary judgments with content bigrams, with data points distinguished by the publisher-assigned genre.

it is not feasible to predict a continuous variable exactly. Instead we report the coefficient of determination (R^2). This metric captures the percentage of variation in the data that the model explains by contrasting the errors of the model predictions with those of the null model which always predicts the mean of the data. R^2 can be contrasted with the root mean squared error, also known as the standard error of the estimate, or the norm of residuals, which measures how close the predictions are to the target on average. In contrast with R^2, this metric has the same scale as the original data, and lower values are better.

The regression scores are shown in Table 4. Predicting the bad/good scores is again more difficult. The regression results for literariness predictions are visualized in Figure 3. Each data point represents a single book. The x-axes show the literariness ratings from survey participants, while the y-axes show the predictions from the model. The diagonal line shows

what the perfect prediction would be, and the further the data points (novels) are from this line, the greater the error. On the sides of the graphs the histograms show the distribution of the literariness scores. Notice that the model based on content bigrams mirrors the bimodal nature of the literariness ratings, while the histogram of predicted literariness scores based on style bigrams shows only a single peak.

Figure 5 shows the same regression results with the publisher-assigned genres highlighted. The graph shows that predicting the literariness of thrillers is more difficult than predicting the literariness of the more literary rated novels. Most thrillers have ratings between 3.4 and 4.3, while the model predicts a wider range of ratings between 3.3 and 5.0; i.e., the model predicts more variation than actually occurs. For the literary novels both the predicted and actual judgments show a wide range between 4.5 and 6.5. The actual judgments of the literary novels are about 0.5 points higher than the predictions. However, there are novels at both ends of this range for which the ratings are well predicted. Judging by the dispersion of actual and predicted ratings of the literary novels compared to the thrillers, the model accounts for more of the variance within the ratings of literary novels.

It should be noted that while in theory 100 % is the perfect score, the practical ceiling is much lower due to the fact that the model is trying to predict an average rating—and because part of the variation in literariness will only be explainable with richer features, text-extrinsic sociological influences, or random variation.

5 Interpretation

As the experiments show, there are textual elements that allow a machine learning model to distinguish between works that are perceived as highly literary as opposed to less literary ones—at least for this dataset and survey. We now take a closer look at the features and predictions of the literary classification task to interpret its success.

5.1 Content

When we look at the forty bigrams that perform best and worst for the literary novels (cf. Table 5), we can identify a few tendencies.

The book, *a book*, *a letter*, and *to write* are also part of the most important features, as well as *the bar*, *a cigarette*, and *the store*. This suggests a certain pre-digital situatedness, as well as a reflection on the writing process. Interestingly enough, in contrast to *the book* and *letter* that are most discriminating, negative indicators contain words related to modern technology: *mobile phone* and *the computer*. Inspection of the novels shows that the literary novels are not necessarily set in the pre-digital age, but that they have fewer markers of recent technology. This might be tied to the adage in literary writing that good writing should be 'timeless'—which in practice means that at the very least a novel should not be too obvious in relating its settings to the current day. It could also show a hint of nostalgia, perhaps connected to a romantic image of the writer.

In the negative features, we find another time-related tendency. The first is indications of time—*little after*, and in Dutch 'tot nu' and 'nu toe', which are part of the phrase 'tot nu toe' (*so far* or *up until now*), *minutes after* and *ten minutes*; another indicator that awareness of time, albeit in a different sense, is not part of the 'literary' discourse. Indicators of location are *the building*, *the garage/car park*, and *the location*, showing a different type of setting than the one described above. We also see indicators of homicide: *the murder*, and *the investigation*. Some markers of colloquial speech are also found in the negative markers: *for god's sake* and *thank you*, which aligns with a finding of Jautze et al (2013), where indicators of colloquial language were found in low-brow literature.

It is possible to argue, that genre is a more important factor in this classification than literary style. However, we state that this is not particular to this research, and in fact unavoidable. The discussion of how tight genre and literariness are connected, has been held for a long time in literary theory and will probably continue for years to come. Although it is not impossible for so called 'genre novels' to gain literary status (cf. Margaret Atwood's scifi(-like) work for instance—although she objects to such a classification; Hoby 2013), it is the case that certain topics and genres are considered to be less literary than others. The fact that the literary novels are apparently not recognised by proxy, but on an internal coherence (cf. section 4), does make an interesting

weight	literary features, content		weight	non-literary features, content	
12.1	*de oorlog*	the war	-6.1	*de moeder*	the mother
8.1	*het bos*	the forest	-5.1	*keek op*	looked up
8.1	*de winter*	the winter	-4.9	*mijn hoofd*	my head
6.6	*de dokter*	the doctor	-4.9	*haar moeder*	her mother
5.8	*zo veel*	so much	-4.7	*mijn ogen*	my eyes
4.8	*nog altijd*	yet still	-4.7	*ze keek*	she looked
4.5	*de meisjes*	the girls	-4.5	*mobiele telefoon*	mobile telephone
4.3	*zijn vader*	his father	-4.2	*de moord*	the murder
4.0	*mijn dochter*	my daughter	-4.0	*even later*	a while later
3.9	*het boek*	the book	-3.8	*nu toe*	(until) now
3.8	*de trein*	the train	-3.5	*zag ze*	she saw
3.7	*hij hem*	he him	-3.4	*ik voel*	I feel
3.7	*naar mij*	at me	-3.3	*mijn man*	my husband
3.5	*zegt dat*	says that	-3.2	*tot haar*	to her
3.5	*het land*	the land	-3.2	*het gebouw*	the building
3.5	*een sigaret*	a cigarette	-3.2	*liep naar*	walked to
3.4	*haar vader*	her father	-3.1	*we weten*	we know
3.4	*een boek*	a book	-3.1	*enige wat*	only thing
3.2	*de winkel*	the shop	-3.1	*en dus*	and so
3.1	*elke keer*	each time	-3.0	*in godsnaam*	in god's name
weight	literary features, style		weight	non-literary features, style	
21.8	*! WW*	! VERB ,	-13.8	*nu toe*	until now
20.5	*u ,*	you (FORMAL) ,	-13.4	*en dus*	and so
18.0	*haar haar*	her her	-13.4	*achter me*	behind me
16.5	*SPEC :*	NAME :	-13.2	*terwijl ik*	while I
15.4	*worden ik*	become I	-13.1	*tot nu*	until now

Table 5: The top 20 most important content features and top 5 most important style features of literary (left), and non-literary texts (right), respectively.

case for the literary novel to be a genre on its own. Computational research into genre differences has proven that there are certain markers that allow for a computer to make an automated distinction between them, but it also shows that interpretation is often complex (Moretti, 2005; Allison et al., 2011; Jautze et al., 2013). Topic modelling might give some more insight into our findings.

5.2 Style

A stronger case against genre determining the classification is the success of the function words in the task. Function words are not directly related to themes or topics, but reflect writing style in a more general sense. Still, the results do not rule out the existence of particular conventions of writing style in genres, but in this case the distinction between literariness and genre becomes more subtle. Function words are hard to interpret manually, but we do see in the top 20 (Table 5 shows the top 5) that the most discriminating features of less literary texts contain more question marks (and thus questions), and more numerals ('TW')—which can possibly be linked to the discriminative qualities of time-indications in the content words. Some features in the less-literary set appear to show more colloquial language again, such as *ik mezelf* ('I myself'), *door naar* ('through/on to'; an example can be found in the sentence *'Heleen liep door naar de keuken.'*, which translates to 'Heleen walked on to the kitchen', a sound grammatical construction in Dutch, but perhaps not a very aesthet-

ically pleasing one). A future close reading of the original texts will give more information on this intuition.

In future work, more kinds of features should be applied to the classification of literature to get more insight. Many aspects could be studied, such as readability, syntax, semantics, discourse relations, and topic coherence. Given a larger data set, the factors genre and translation/original can be controlled for.

The general question which needs to be answered is whether a literary interpretation of a computational model is even possible. The material to work with (the features), consist of concise sets of words or even part-of-speech tags, which are not easy to interpret manually; and they paint only a small part of the picture. The workings of the machine learning model remain largely hidden to the interpreter. This is an instance of the more general problem of the interpretability of results in computational humanities (Bod, 2013). In the specific case of literature, we can observe that readers of literature follow a similar pattern: literature can be recognized and appreciated, but it is hard to explain what makes texts literary, let alone to compose a highly literary work.

5.3 Good and bad predictions

In Figure 5, we can see both outliers and novels that are well predicted by the regression model. Here we discuss a few and suggest why the model does or does not account for their perceived literariness.

Emma Donoghue - Room A literary novel that is rated as highly literary (5.5), but with a lower prediction (3.8). This may be because this novel is written from the perspective of a child, with a correspondingly limited vocabulary.

Elizabeth Gilbert - Eat, Pray Love A novel with a low literariness rating (3.5), but a high prediction (5.2) by the model. This novel may be rated lower due to the perception that it is a novel for women, dealing with new age themes, giving it a more specific audience than the other novels in the dataset.

Charles Lewinsky - Melnitz A novel that is both rated (5.7) and predicted (5.7) as highly literary. This novel chronicles the history of a Jewish family including the events of the second world war. This subject, and the plain writing style makes it stand out from the other novels.

Erwin Mortier - While the Gods Were Sleeping The most highly rated (6.6) literary novel in the dataset, with a high prediction (5.7). A striking feature of this novel is that it consists of short paragraphs and short, often single line sentences. It features a lot of metaphors, analogies, and generally a poetic writing style. This novel also deals with war, but the writing style contrasts with Lewinsky, which may explain why the model's prediction is not as close for this novel.

6 Related Work

Previous work on classification of literature has focused on authorship attribution (e.g., Hoover, 2003; van Cranenburgh, 2012) and popularity (Ashok et al., 2013). The model of Ashok et al. (2013) classifies novels from Project Gutenberg as being successful or not using stylometric features, where success is based on their download counts. Since many of the most downloaded novels are classics, their results indirectly relate to literariness. However, in our data set all texts are among the most popular books in a fixed time span (cf. section 2), whereas the less successful novels in their data set differ much more in popularity from the successful novels. To the best of our knowledge, our work is the first to directly predict the literariness of texts in a computational model.

There is also work on the classification of the quality of non-fiction texts. Bergsma et al. (2012) work on scientific articles with a similar approach to ours, but including syntactic features in addition to bag-of-words features. Louis and Nenkova (2013) present results on science journalism by modelling what makes articles interesting and well-written.

Salganik et al. (2006) present an experimental study on the popularity of music. They created an artificial "music market" to study the relationship between quality and success of music, with or without social influence as a factor. They found that social influence increases the unpredictability of popularity in relation to quality. A similar effect likely plays a role in the reader judgments of the survey.

7 Conclusion

Our experiments have shown that literary novels share significant commonalities, as evidenced by the performance of machine learning models. It is still a challenge to understand what these literary commonalities consist of, since a large number of word features interact in our models. General quality is harder to predict than literariness.

Features related to genre (e.g., *the war* in literary novels and *the homicide* in thrillers) indicate that genre is a possible confounding factor in the classification, but we find evidence against the notion that the results are solely due to genre. One aspect that stood out in our analysis of content features, which is not necessarily restricted to genre (or which might indicate that the literary novel is a genre in and of itself), is that setting of space and time rank high among the discriminating features. This might be indicative of a 'timeless quality' that is expected of highly literary works (where words as *book* and *letter* are discriminative)—as opposed to more contemporary settings in less literary novels (*computer* and *mobile phone*). Further study is needed to get more insight into these themes and to what extent these are related to genre differences or a literary writing style.

The good performance of style features shows the importance of writing style and indicates that the classification is not purely based on topics and themes. Although genres may also have particular writing styles and thus associated style features, the fact that good results are obtained with two complementary feature sets suggests that the relation between literariness and text features is robust.

Finally, the regression on content and function words shows that the model accounts for more than just genre distinctions. The predictions within genres are good enough to show that it is possible to distinguish highly literary works from less literary works. This is a result that merits further investigation.

Acknowledgments

We are grateful to Karina van Dalen-Oskam, Rens Bod, and Kim Jautze for commenting on drafts, and to the anonymous reviewers for useful feedback. 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.

References

Sarah Danielle Allison, Ryan Heuser, Matthew Lee Jockers, Franco Moretti, and Michael Witmore. 2011. Quantitative formalism: an experiment. Stanford Literary Lab pamphlet. http://litlab.stanford.edu/LiteraryLabPamphlet1.pdf.

Vikas Ashok, Song Feng, and Yejin Choi. 2013. Success with style: Using writing style to predict the success of novels. In *Proceedings of EMNLP*, pages 1753–1764. http://aclweb.org/anthology/D13-1181.

Shane Bergsma, Matt Post, and David Yarowsky. 2012. Stylometric analysis of scientific articles. In *Proceedings of NAACL*, pages 327–337. http://aclweb.org/anthology/N12-1033.

Rens Bod. 2013. Who's afraid of patterns?: The particular versus the universal and the meaning of humanities 3.0. *BMGN – Low Countries Historical Review*, 128(4). http://www.bmgn-lchr.nl/index.php/bmgn/article/view/9351/9785.

Pierre Bourdieu. 1996. *The rules of art: Genesis and structure of the literary field*. Stanford University Press.

Hermione Hoby. 2013. Margaret Atwood: interview. The Telegraph, Aug 18. http://www.telegraph.co.uk/culture/books/10246937/Margaret-Atwood-interview.html.

David L. Hoover. 2003. Frequent collocations and authorial style. *Literary and Linguistic Computing*, 18(3):261–286. http://llc.oxfordjournals.org/content/18/3/261.abstract.

Kim Jautze, Corina Koolen, Andreas van Cranenburgh, and Hayco de Jong. 2013. From high heels to weed attics: a syntactic investigation of chick lit and literature. In *Proc. of workshop Computational Linguistics for Literature*, pages 72–81. http://aclweb.org/anthology/W13-1410.

Annie Louis and Ani Nenkova. 2013. What makes writing great? first experiments on article quality prediction in the science journalism domain. *Transactions of the Association for Computational Linguistics*, 1:341–352. http://aclweb.org/anthology/Q13-1028.

Franco Moretti. 2005. *Graphs, maps, trees: abstract models for a literary history*. Verso.

Fabian Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Matthew J. Salganik, Peter Sheridan Dodds, and Duncan J. Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–856.

Andreas van Cranenburgh. 2012. Literary authorship attribution with phrase-structure fragments. In *Proceedings of CLFL*, pages 59–63. Revised version: `http://andreasvc.github.io/clfl2012.pdf`.