



UvA-DARE (Digital Academic Repository)

Topic modeling literary quality

Jautze, K.; van Cranenburgh, A.; Koolen, C.

Publication date

2016

Document Version

Final published version

Published in

Digital Humanities 2016

[Link to publication](#)

Citation for published version (APA):

Jautze, K., van Cranenburgh, A., & Koolen, C. (2016). Topic modeling literary quality. In W. Eder, & J. Rybicki (Eds.), *Digital Humanities 2016: Concurrence abstracts : Jagiellonian University & Pedagogical University, Kraków, 11-16 July 2016* (pp. 233-237). European Association for Digital Humanities [etc.]. <http://dh2016.adho.org/abstracts/95>

General rights

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), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

The European Association for Digital Humanities (EADH)
Association for Computers and the Humanities (ACH)
Canadian Society for Digital Humanities / Société canadienne des humanités numériques (CSDH/SCHN)
centerNet
Australasian Association for Digital Humanities (aaDH)
Japanese Association for Digital Humanities (JADH)

Digital Humanities 2016

Conference Abstracts

Jagiellonian University
&
Pedagogical University

Kraków
11–16 July 2016



Kraków 2016

Topic Modeling Literary Quality

Kim Jautze

kim.jautze@huygens.knaw.nl

Huygens ING, Royal Netherlands Academy of Arts and Sciences, Netherlands

Andreas van Cranenburgh

andreas.van.cranenburgh@huygens.knaw.nl

Huygens ING, Royal Netherlands Academy of Arts and Sciences, Netherlands; Institute for Logic, Language and Computation, University of Amsterdam

Corina Koolen

c.w.koolen@uva.nl

Institute for Logic, Language and Computation, University of Amsterdam

Introduction

To what extent can topic models explain variation in perceptions of literary quality? We try to find correlations between topics and judgments of literary quality using a topic model of 401 recent bestselling Dutch novels. Instead of examining topics on a macro-scale in a geographical or historical interpretation (e.g., Jockers 2013; Riddell 2014), we take a new perspective: whether novels have a dominant topic in their topic distributions (mono-topicality), and whether certain topics may express an explicit or implicit genre in the corpus. We hypothesize that there is a relationship between these aspects of the topic distributions and perceptions of literary quality. We then interpret the model by taking a closer look at the topics in a selection of the novels.

Riddle survey and corpus

This research is part of a Dutch computational humanities project called The Riddle of Literary Quality. In the project we aim to identify textual features that may play a role in readers' evaluations of a novel as being good or bad and as high or low literature. We analyze a corpus of 401 contemporary Dutch-language (including translated) novels in search of textual features they have in common. Within our corpus there is a small variety of novelistic genres, which can be roughly divided into suspense, romantic and general novels. The readers' judgments were gathered in a large online survey. We asked a general public to rate the novels they had read on a 7-point scale from *definitely not* through *highly literary*. Approximately 14,000 respondents participated, providing us with much data on the perceived quality of our 401 novels.¹ The mean rating over all 401 novels is 4.2, with 2.1 being the lowest rating for *Fifty Shades of Grey* by E.L. James, and 6.6 the highest for Julian Barnes' *The Sense of an Ending*.

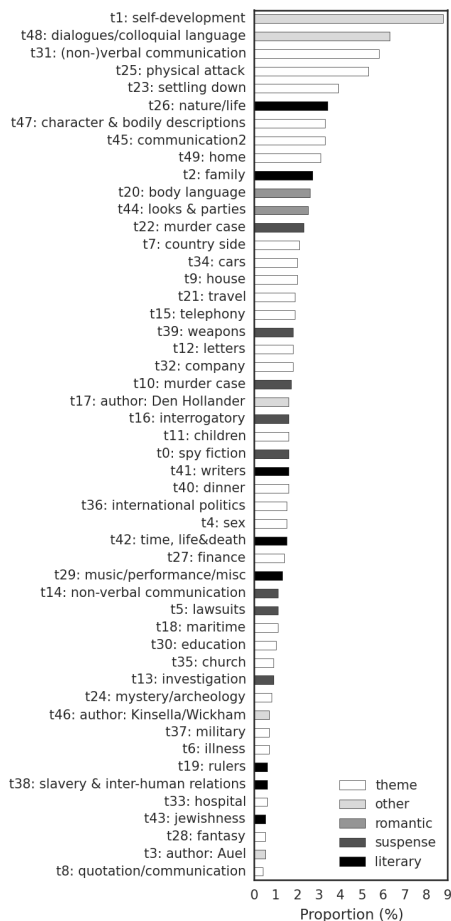


Figure 1: Overview of topics, sorted by proportion of the corpus

Topic model

A topic model aims to automatically discover topics in a collection of documents. We use Latent Dirichlet Allocation (Blei et al., 2003), which assumes the documents have been generated from a fixed number of probability distributions (the topics) over words. The topics reflect word co-occurrence patterns. We preprocess the novels by lemmatizing the words, removing punctuation, function words and names, and splitting the remaining text in chunks of 1000 tokens. We use MALLET to create a topic model with 50 topics. Fig. 1 shows an overview of the topics with their proportion across the corpus.

We have attempted to identify topics for novels with high literary ratings, and topics specific for suspense and romantic novels. According to Jockers and Mimno (2013), the topics can be used to identify literary themes. They use the terms “theme” and “topic” as “proxies for [...] a type of literary content that is semantically unified and recurs with some degree of frequency or regularity throughout and across a corpus” (p. 751). We found that three topics are specific to a single author (for instance t3), and about a third seem genre specific. By inspecting the most important words for each topic we found that most topics

(genre related or not) indeed cohere with certain themes (cf. Fig.1). This suggests that the choice for 50 topics is neither too small nor too high.

Quantitative analysis

We aim to gain insight into the distribution of topics in relation to the literary ratings of the novels (predicting literary ratings is not the main aim here). In order to interpret the topic distributions, we introduce the concept of mono-topicality.² A mono-topical novel contains little diversity in topic distribution, which means that one or two topics are dominant throughout the novel. A novel which shows more variation in topics has a more even distribution of topics, i.e., such a novel has a larger topic diversity. Fig. 2 shows an example of both cases.

The x-axis shows the distribution of topics, sorted from least to most prevalent. In John Grisham’s *The Appeal*, topic 5 (“lawsuits”) has a proportion of 47.8 % of all 50 topics. This novel is more mono-topical than the Franzen’s *Corrections*, which has a more balanced distribution of topic proportions.

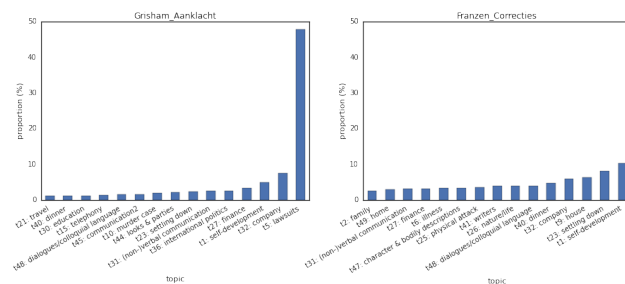


Figure 2: Distribution of the top 15 topics in novels with high (left) and low (right) mono-topicality

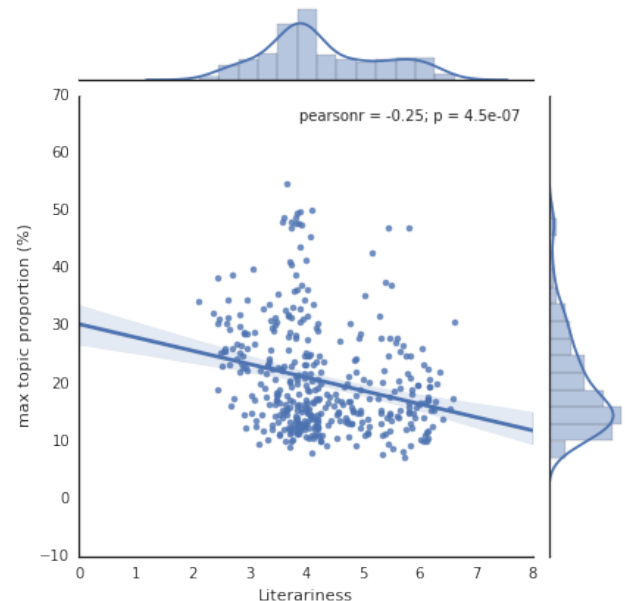


Figure 3: Correlation between share of the most prominent topic per book and mean literariness ratings

Author_Title	Survey rating	% topic 29	Relationships	Friend / family	Parent-child	love	ri-valry	illness & Heath	Ar-tistic pro-fes-sion	writer / edi-tor	Roth-er	True story	auto-bio-graphic	his-tory
Hart_Verlovingstijd	5.9	24.8	X				X							
Rosenboom_ZoeteMond	6.2	22.5				X	X							
Lanoye_Sprake-loos	6.4	21.8			X			X	X				X	
Dewulf_KleineD-agen	6.0	19.6	X						X				X	
Rosenboom_Me-chanica	6.2	17.0							X					
Heijden_Tonio	6.3	16.8			X			X	X				X	
Verhulst_Laatste-LiefdeVan	5.8	16.5			X		X							
Lanoye_Heldere-Hemel	5.8	15.7	X				X							X
Springer_Quad-riga	6.0	15.7				X			X					X
Mortier_Ges-tameldLiedboek	6.5	15.5			X			X	X				X	
Kooten_Verreki-jker	5.0	15.2			X				X				X	
Moor_SchilderE	5.9	14.7				X				X				X
Meisje														
Zwagerman_Duel	5.5	14.6									X			
Giphart_IJsland	5.3	13.0			X						X			
Dorrestein_Stief-moeder	5.5	8.7	X		X									

Table 1: Themes in fifteen highly literary novels, all of which are originally Dutch Table 2: Six topics from the model that address themes present in the fifteen highly literary novels, but which are not the most prominent as topics in those novels

We hypothesize that the less mono-topical a novel is, the higher the perceived literariness by readers will be. And indeed, Fig. 3 shows that there is a statistically significant correlation between the diversity of topics of a book and its perceived literariness. Books with a single, highly prominent topic, such as Grisham’s, tend to be seen as less literary.

Interpretation

There are several possible explanations for the correlation. Genre novels could have a tendency to single out certain topics, as they deal with more ‘typical’ or genre-specific subject matter than do general novels. If this were the case, we would simply be finding that genre novels are

considered to be less literary than general novels, and this would tell us little about literary quality in a more general sense. General novels in the other hand, deal with all sorts of subjects and themes, across and beyond 'genre' subjects, and therefore a topic model may not be able to single out thematic groups of words common to these novels, and thus may not find one single prominent topic. A third explanation could be that highly literary novels *do* deal with specific themes and subjects which are also part of genre novels, but that these are described in wordings that are more implicit or subtle, and therefore do not come up as single, clear topics. If this were the case, that would bring us closer to an explanation of what topics have to do with literary quality. These explanations are not mutually exclusive and we will explore the topic model here to examine the value of the second and third explanation.

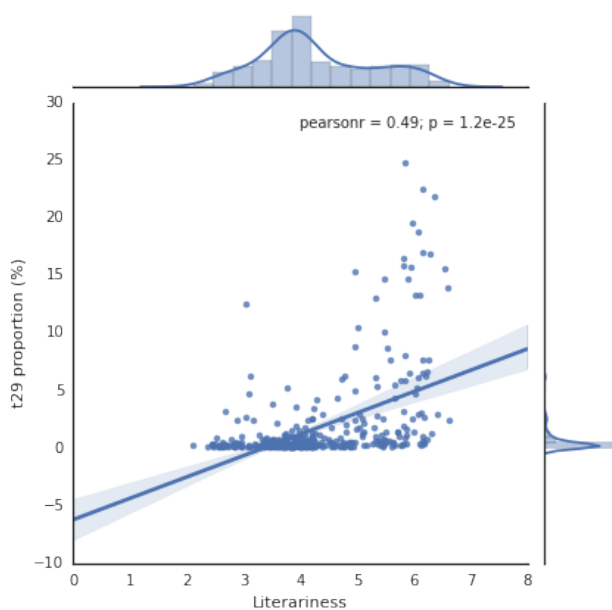


Figure 4: Correlation between topic 29 proportion and mean literariness ratings

The topic that shows the highest correlation ($r=0.49$) with literary appreciation is topic 29; cf. Fig. 4. This topic is most prominent in fifteen originally Dutch general novels. The twenty words in topic 29 with the highest weights are *begin, music, play, occasion, first, the first, sing, only, year, one, stay, sometimes, even, new, own, always, high, exact(ly), bike, appear*. They show little coherence, making it hard to interpret their context, although 'music and performance' appears to be part of it. To find out more about the novels in which this topic is prominent, we consult a Dutch website maintained by librarians called *Literatuurplein*, which provides information on the themes and content of Dutch novels.

Most of these novels show similarities in themes, such as family relationships. In ten of the novels the protagonist has an artistic profession: a couple of writers, a painter and a stand-up comedian. None of them has a musical or

acting career, despite the 'music and performance' words; and vice versa, none of the twenty most prominent words concern writing.

All in all, at first glance topic 29 seems not to address the themes and content of the novels, whereas most other topics in the model do concern specific themes (cf. Fig. 1 and Table 2).

Topic	Name	Top 10 words with highest weight
2	Family relations I	father, mother, child, year, son, girl, brother, woman, older, daughter
6	Health I	doctor, body, pain, illness, pill, blood, death, medicine, child, patient
11	Family relations II	child, mother, mom, baby, dad, little, cry, hand, time, grandma
12	Writing & memories	picture, letter, write, read, paper, book, year, day, envelope, memories
33	Health II	doctor, hospital, patient, women, bed, lie, nurse, room, hall, hour
41	Novels	book, writing, story, year, word, writer, human, novel, time

Table 2: Six topics from the model that address themes present in the fifteen highly literary novels, but which are not the most prominent as topics in those novels

For instance, topic 2 and 11 address family relations, topic 12 and 41 are about writing novels, and topic 6 and 33 concern health issues. These topics are present, but as smaller topics. This shows that the second explanation, of the general novels not sharing themes, is not valid. It could be an indication though that the highly literary novels indeed use a more subtle way of describing themes similar to other novels in our corpus, our third explanation. As a final note, in topic 29 there are proportionally more adverbs than in the other topics mentioned, which contain more nouns. Perhaps this shows that style is a more shared element in literary novels than the choice of words. In other words, this brief analysis shows that there is merit to our third explanation. This will therefore become a new hypothesis for further research.

Conclusion

We have explored a topic model of contemporary novels in relation to genre and literariness, and shown that topic diversity correlates with literary ratings. Most topics express a clear theme or genre. However, topic 29, the most literary topic, does not. It rather appears to be associated with a particular Dutch literary writing style.

Bibliography

Algee-Hewitt, M., Heuser R., and Moretti, F. (2015). On paragraphs. Scale, themes, and narrative form. Stanford Literary

Lab pamphlet 10. <http://litlab.stanford.edu/LiteraryLabPamphlet10.pdf>

- Blei, D. M., Ng, A. Y., and Jordan, M. I.** (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.
- Jockers, M. L. and Mimno, D.** (2013). Significant Themes in 19th-Century Literature. *Poetics* 41(6):750–69. <http://dx.doi.org/10.1016/j.poetic.2013.08.005>
- Jockers, M. L.** (2013). *Macroanalysis: Digital methods and literary history*. University of Illinois Press.
- Riddell, Allen** (2014). How to read 22,198 journal articles: Studying the history of German studies with topic models. In Erlin, M. and Tatlock, L. (eds), *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, Rochester, New York: Camden House, pp. 91–114.

Notes

- ¹ Extensive details on the survey will be published in two articles, one of which is in submission.
- ² After submitting this abstract, we discovered the Literary Lab pamphlet by Algee-Hewitt et al. (2015), who independently devised a concept called mono-topicality.