What Do We Talk About When We Talk About Topic?

Joris J. van Zundert[1,2], Marijn Koolen[1,2], Julia Neugarten[3], Peter Boot[1], Willem van Hage[4] and Ole Mussmann[4]

[1]KNAW Huygens Institute, Amsterdam, the Netherlands
[2]DHLab, KNAW Humanities Cluster, Amsterdam, the Netherlands
[3]Radboud University Nijmegen, Nijmegen, the Netherlands
[4]eScience Center, Amsterdam, the Netherlands

Abstract

We apply Top2Vec to a corpus of 10,921 novels in the Dutch language. For the purposes of our research we want to understand if our topic model may serve as a proxy for genre. We find that topics are extremely narrowly related to an existing genre classification historically created by publishers. Interestingly we also find that, notwithstanding careful vocabulary filtering as suggested by prior research, various other signals, such as author signal, stubbornly remain.

Keywords

literary fiction, novels, computational literary studies, topic models, top2vec,

1. Introduction

This short paper presents the preliminary results of topic-modelling 10,000+ contemporary novels in the Dutch language, published between 2009 and 2019. The purpose of this paper is to understand how topics yielded by topic modelling relate to genre in this corpus. This is an important step for the ultimate aim of the project that this paper relates to, which is to understand how reader impact is distributed across genre and topic.

While the results of topic models are often taken at face value as semantically meaningful, such assumptions risk mistaking artefacts of a specific corpus and its structure for content-level literary topics. Only a small part of the topics resulting from our analysis relate to content words that share clearly recognizable subject matter (including words relating to football, medicine or music). Many other topics group together language use, historical period, or geographical location. In literary studies, such groupings do not constitute the topic or theme of a novel. We conclude that contextual information from novels in a corpus, such as the location, historical period, or language community in which they are situated, and even author signal co-shape topics. Additionally, we find that the topics we find are very strongly correlated to

CHR 2022: Computational Humanities Research Conference, December 12–14, 2022, Antwerp, Belgium

joris.van.zundert@huygens.knaw.nl (J. J. v. Zundert); marijn.koolen@gmail.com (M. Koolen); julia.neugarten@ru.nl (J. Neugarten); peter.boot@huygens.knaw.nl (P. Boot); w.vanhage@esciencecenter.nl (W. v. Hage); o.mussmann@esciencecenter.nl (O. Mussmann)
https://jorisvanzundert.net/ (J. J. v. Zundert); https://marijnkoolen.com/ (M. Koolen)
0000-0003-3862-7602 (J. J. v. Zundert); 0000-0002-0301-2029 (M. Koolen); 0000-0003-3314-9445 (J. Neugarten); 0000-0002-7399-3539 (P. Boot); 0000-0002-6478-3003 (W. v. Hage); 0000-0002-3803-4287 (O. Mussmann)

genres, like horror, romance, crime, etc. Our analysis invites reflection on the usefulness of topic modelling as a tool for computational literary studies (CLS).

Because of copyright, we cannot share the texts of the novels on which the topic models are based. However, we can and will eventually share the intermediate results and our code through the GitHub repository of the project.[1]

2. Topic modelling and literature

Various technologies have been used in CLS and other fields to establish the semantic value of topics. In CLS, for example, researchers have used frequent words [20], (some variety of) keywords [3, 19], top-down methods such as the UCREL Semantic Analysis System [18], and systematic manual procedures [1].

A commonly applied technique is topic modelling, which algorithmically identifies groups of words that tend to co-occur in a large collection of documents [10]. Although introduced in a non-fiction context, this has been applied to fiction (e.g. [6, 14, 22, 15, 12, 7, 11, 25]). Topic modelling has also been applied to the discourse of literary studies [9] and to online literature reviews [28].

Some previous research in CLS suggests that topic modelling literary works leads to semantically cohesive topics and an overarching understanding of the subject matter represented in those works [2]. In *Macroanalysis* [13], Jockers used "Themes" as the title for his chapter about topic modelling. This title suggests that topic modelling is a technique for unearthing "the" theme of a novel, viz. "a salient abstract idea that emerges from a literary work's treatment of its subject-matter" (the first meaning of "theme" defined in [5]). Schröter and Du [23] suggest that "sujet" is similar to literary topic. Lundy [15] uses topic modelling on a corpus of ca. 1,000 recent popular U.S. novels. He creates one set of general topics, analyzing entire novels at a time. Additionally, he creates a set of more specific topics using individual sentences from these books. The focus of the study is on how these topics are distributed over genres. Jautze et al. [12] topic model 400 recent Dutch works of fiction, and use the resulting topics to predict a reader response variable (perceived literariness). Others (e.g. [11]) argue that topic models do not generate semantically coherent descriptions of topics meaningful to CLS.

Of particular interest here are [22], where it is argued that "strong genre signals exist [...] on the levels of function words, content words and syntactic structure" but that they "also exist on the level of theme or topic"; [25] arguing that topics relate to structural rather than content elements of text; and [24] where it is shown that topics strongly correlate to meta-textual features such as author and genre.

3. Problem

There is no hard scientific consensus on what textual features constitute genre, and as, for instance, [29] argues there may be good reason to question canonical genre classifications. Topics from topic models on the other hand are notoriously hard to clearly relate to literary

[1] https://github.com/impact-and-fiction

constructs or categories [22, 25, 24]. If we are interested in the relation between quantifiable textual features and readers' preferences, our question becomes how bottom up topics from a topic model relate to given genre metadata.

We operationalise genre categories using Dutch NUR-codes (Nederlandse Uniforme Rubrieksindeling or Dutch Uniform Categories classification). These were introduced in 2002 as a market monitoring instrument and succeed the comparable NUGI codes that have been in use for the same reason since 1987. NUR is a practical marketing instrument devised and applied by publishers. It is largely ignored in Dutch literary culture, and it goes largely unnoticed by readers who mostly encounter it because bookshops tend to sort and arrange their product range according to the system [26, 27]. NUR can be regarded as a rough approximation of the concept of genre as it is understood by booksellers and readers.

The examination of the correlation between genre (NUR) and topic can be broken down into several sub-questions. How is a topic related to specific NUR codes? What is the topic distribution of a NUR code? What is the NUR distribution of the books most associated with a topic? As a first step, we determine how strongly topics are associated with NUR codes. Ideally, such associations are unrelated to topics that emerge from our model as the artefacts of corpus features like author signal, translation, and corpus composition as these are irrelevant with regard to canonical genre categories.

We make the results of our topic modelling concrete and useful to CLS in two ways: by examining the correlations between the topics detected in our corpus and one possible operationalisation of genre, and by reflecting on the usefulness of topic modelling for literary analysis in light of our results.

4. Method

4.1. Data

Courtesy of an agreement with the Dutch national library and seven Dutch publishing houses (representing multiple publishers) we have access to the full text of 10,921 Dutch-language novels, published between 2009 and 2019 in the Netherlands. Table 1 lists a number of general statistics for the corpus used in this research.

Table 1
Corpus statistics

		Per book				
Element	Number	Min	Max	Median	Mean	Std.dev
Novels	10,921	1	1	1	1	1
$Windows_{5000}$	153,553	1	104	14.1	11.0	11.3
Paragraphs	24,356,023	1	21355	2230.2	1668.0	2000.9
Sentences	104,511,706	1	80140	9569.8	7213.0	8322.8
Words	931,220,543	1	655744	85268.8	66577.0	71191.9

The collection is based on the EPUBs deposited by publishers at the national library. Therefore, it is a subset of all books published in the Netherlands in this period. The collection is skewed towards more recent books (see Figure 2), primarily because more books are now being made available as ebooks, not because of an increase in publications. The corpus consists of both Dutch novels and at least 2,199 translated novels, mainly from English, German, French and various Scandinavian languages, as well as smaller representations of languages such as Spanish, Italian, and Japanese.

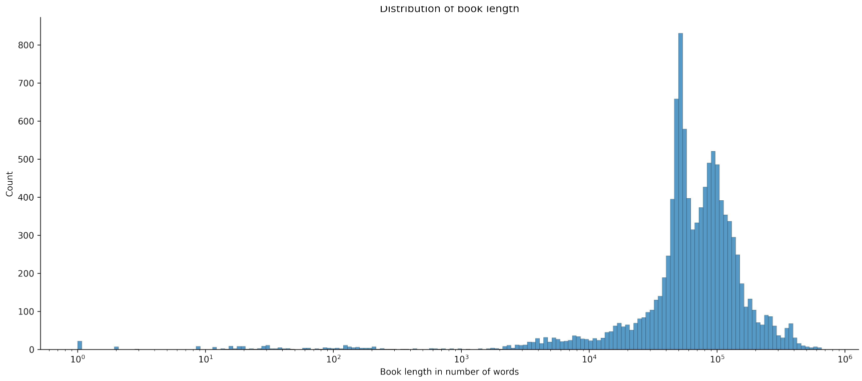

Figure 1: Distribution of book length in number of words for 10,921 Dutch novels

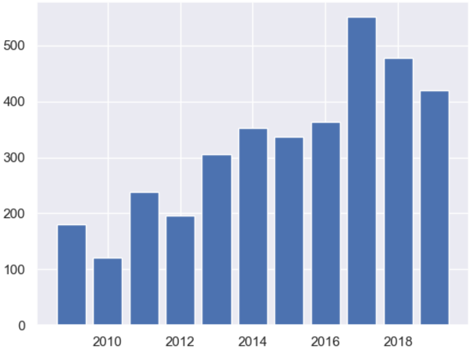

Figure 2: Number of books in the corpus by publication year.

Figure 1 shows the distribution of book lengths in number of words per book. There is one sharp peak around 50,000 words and a lower, less pronounced peak around 90,000 words, roughly coinciding with the conventional lengths of novellas (~80-120 pages) and “full” novels (~300-500 pages). A small number of books are very short. Of the 10,921 novels, 216 novels (2%) are shorter than 1,000 words, and 463 novels (4%) are between 1,000 and 10,000 words. Some of these may be picture books, children’s books, or collections of poetry. Some unusually short books are regular-sized novels for which the text extraction step did not work properly. These are mostly EPUB “incunables”. Apparently publishers needed to get used to the EPUB format and so early EPUBs often have poor file and content structure.

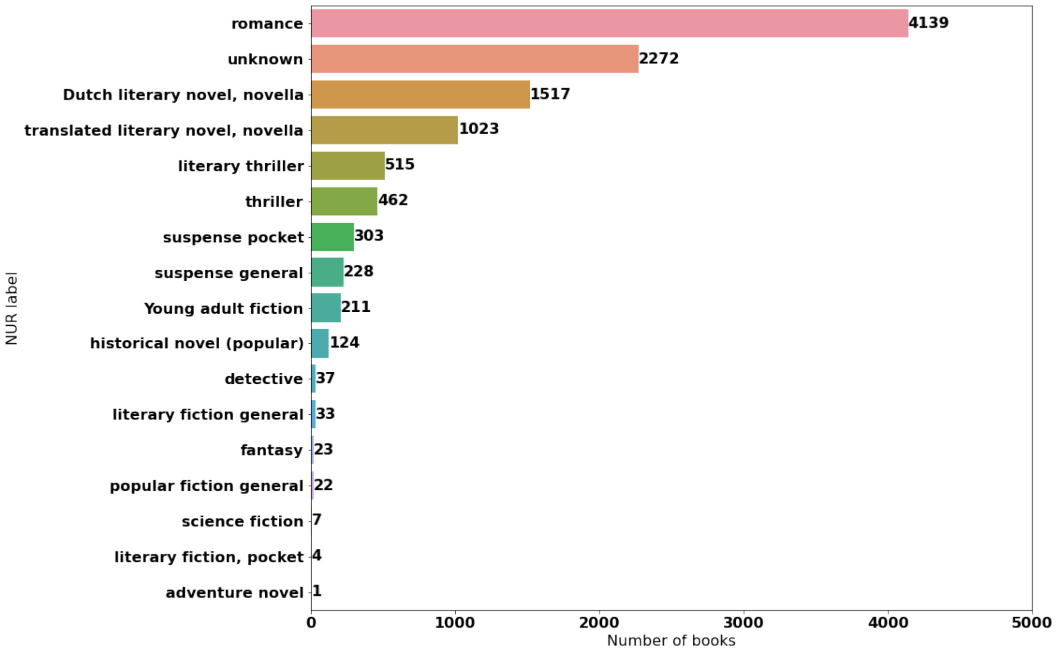

Figure 3: Distribution of books over NUR genre classes

4.2. Preprocessing

Based on existing research [12, 25], we pre-process the novel texts to remove person names and use lemmas rather than full word forms. We tokenise and parse all novels using SpaCy 3.3[2] and remove all word tokens that are part of person entities. SpaCy inserts underscores in lemmas for certain compound words, but sometimes fails to split the words correctly. E.g. for the Dutch word "boekhandel" (en: *book shop*), the SpaCy lemma is almost always "boek_handel", but sometimes SpaCy assigns the lemma "boekh_andel". Therefore, we post-process the lemmas by removing underscores. Based on a sample of the 1,000 most common lemmas containing underscores and the variants containing underscore in a different position, we found that removing the underscores rarely conflates words with different meanings.

We remove common words based on their document frequency, as such words tend to appear in the majority of topics and thus have no discriminating effect between topics. We also remove lemmas that occur in very few books. These tend to include many specific names (many character names are not recognised by SpaCy as person named entities), and very rare vocabulary used by only one or a few authors. This means that we remove lemmas that occur in fewer than 1% or more then 10% of books (fewer than 103 books or more than 1,030 books). This leaves a vocabulary of 36,927 lemmas.

This setup represents a trade-off between corpus coverage and the ability to generate differentiating topics. For future evaluation we aim to vary this bandwidth to gauge the stability of

[2] See https://spacy.io

the topic models inferred (see also 6).

4.3. Segmentation

Next, we need to choose a unit of measure. Based on prior research We test two different unit sizes: the whole novel as a document, and documents constructed from joining a sequence of paragraphs in a novel into segments containing at least 5,000 words. Using whole documents yields rather few topics (95) to relate to 731 existing NUR codes, while the latter choice results in more and smaller documents, and 1,182 topics. Further details and analysis of these different unit are discussed in Appendix A.

4.4. Topic modelling

Given the number and size of documents (or segments) in a corpus, it is difficult to decide the minimum number of relevant topics. [15] used LDA[3] on a set of 1,136 novels and chose 60 as the optimal number of topics based on the fraction of topics they could meaningfully interpret. [12] used LDA on lemmatised 1,000-word segments of 401 Dutch novels and set the number of topics to 50. [25] segmented novels into 300 to 500 word segments, and used a pointwise mutual information (PMI) and a cosine based coherence measure to observe that “as a rule of thumb [...] the number of topics should lie between 100 and 150” [25, p.66].

In this research we applied Top2Vec [4], which in recent studies has compared favorably to LDA and other techniques [21], [8]. Models generated using LDA or PLSA[4] generate a pre-determined number topics as distributions of words. This often means that topically uninformative words have high probabilities in the topics, since they make up a large proportion of all text. In Top2Vec, joint document and word embeddings are learned in a 300-dimensional space, which is projected onto a low-dimensional space using UMAP [17], after which HDBSCAN [16] is used to detect dense document clusters, which determine the number of topics. This ensures that the words nearest a topic vector best describe the topic and its surrounding documents [4, p.3].

5. Results

In this first analysis we use the full corpus, lemmatise tokens, drop persons names (PER according to SpaCy), and cull lemmas appearing in less than 1% or over 10% of all novels. Top2Vec was used in its ’fastlearn’ mode with 8 workers and a standard multilingual universal sentence encoder.[5] When using full novels as unit of measure, this resulted in 95 topics. For 5,000 token windows, Top2Vec generated 1,182 topics.

Labeling the topics that result from topic modelling is a complicated and subjective process which we believe requires annotation and an assessment of inter-annotator agreement. Such

[3] Latent Dirichlet allocation, a common approach to topic modelling, see https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

[4] See https://en.wikipedia.org/wiki/Probabilistic_latent_semantic_analysis

[5] https://tfhub.dev/google/universal-sentence-encoder-multilingual/3

assessment falls outside the scope of the current paper. For this reason, we have chosen to number topics here, rather than give them a semantically meaningful label.

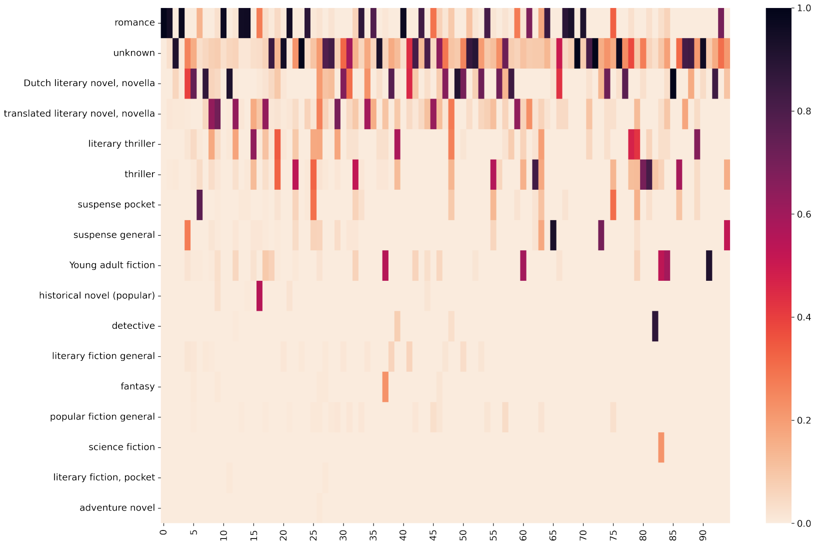

Figure 4: Heatmap of books most associated with a topic (x-axis) and their distribution across NUR codes (y-axis), given entire novels as document unit for topic modelling

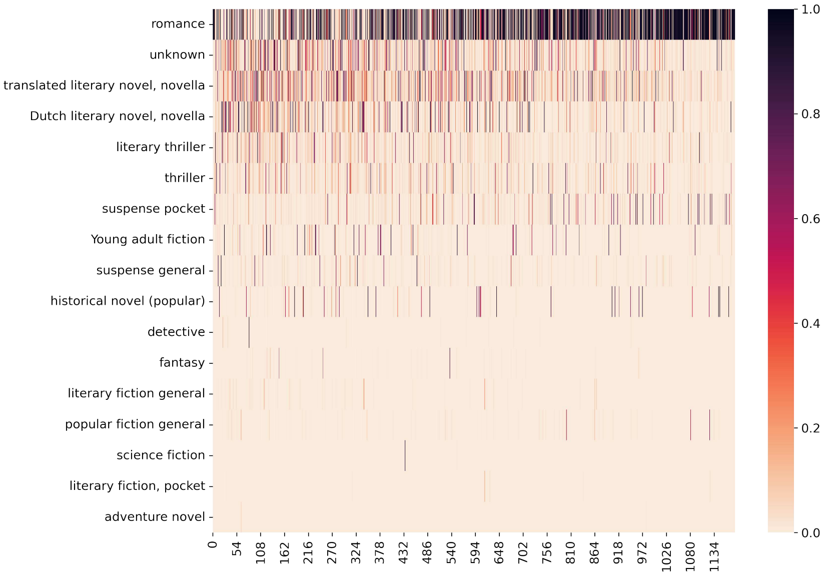

Figure 5: Heatmap of books most associated with a topic (x-axis) and their distribution across NUR codes (y-axis), given text segments of 5,000 tokens for topic modelling

Each book has a nearest topic neighbour, which Top2Vec uses to determine the size of a topic. This way, each topic has a size, expressed as the number of books for which that topic is the nearest neighbour. This allows us to investigate how strongly topics are associated with the NUR codes: we look at the distribution of NUR codes as the percentage of a topic's most associated books that are assigned that NUR code. For instance, the first topic is the nearest topic for 1,655 of the novels, and of these 1,643 (99%) are labelled with NUR code 343 *Romance*. This topic is therefore strongly associated with a single NUR code. We investigate the association of topics and NUR code using a heatmap (see figure 4). The darker a cell, the stronger the topic (x-axis) is associated with a genre (y-axis). Many topics strongly associate with one or two NUR codes. One observation we derive from this is that topic as inferred by Top2Vec is strongly associated with publishers' choices for NUR genre code. This observation can be further corroborated if we use UMAP to create a visualisation of topic clusters where we color each vector according to the NUR it most closely associates with (see figure 6).

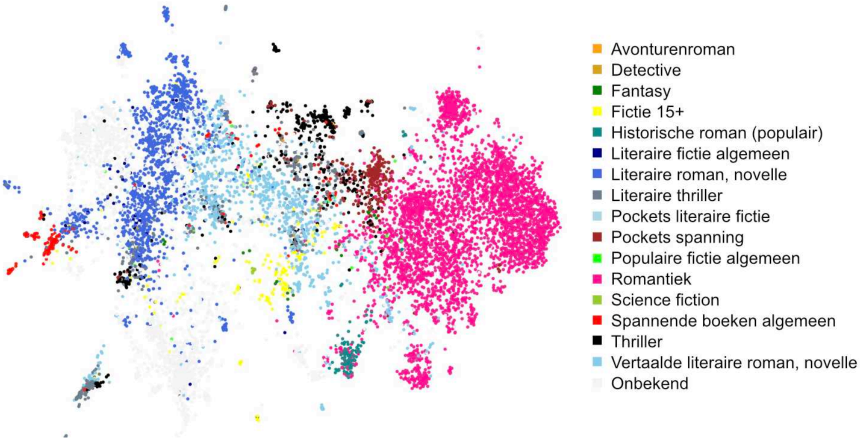

Figure 6: UMAP reduction of topic clusters, topics colored by NUR.

As we have seen in the previous section, the number of topics increases if the unit of measure decreases. If we topic model at the level of whole novels we find only 95 topics. With 5,000 word segments, the resulting model has 1,182 topics, and the association between topic and NUR code is even stronger (see Figure 5). Both the size of the segments (roughly comparable to the size of chapters) and the number of topics found, coincide more closely with intuitions about how literary topics may function. That is, from a literary criticism point of view we would expect topics to be bound more narrowly to chapters or paragraphs rather than to a book as a whole, while 95 topics would seem a tiny set of topics to cover a corpus of over 10,000 novels.

6. Discussion

What becomes clear from the results as depicted in Figures 4 and 5, is that topics generated by Top2Vec are extremely narrowly related to genre, with many topics almost exclusively related to one genre. Furthermore, topics that are strongly associated with the genres "Dutch literary novel" and "Translated literary novel" turn out to contain a large number of geographical

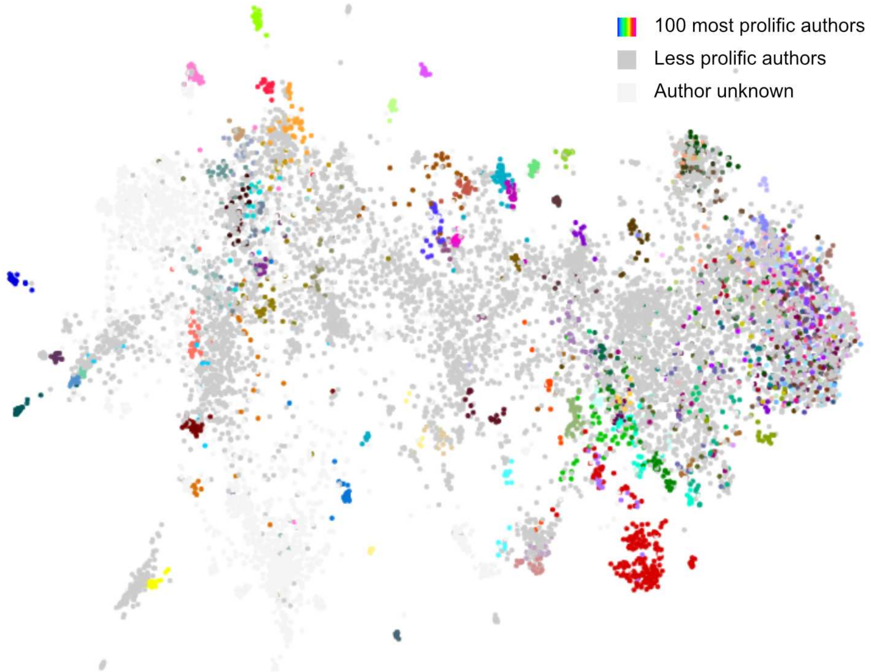

Figure 7: UMAP reduction of topic clusters, topics colored by 100 most prolific authors, others in grey.

indicators (cf. appendix B).

Our results confirm findings from [22, 25, 24]. In all this means that topics as generated by Top2Vec across our corpus will be an adequate proxy for genre in the course of our project. Thus our current result can be summarised as "when we talk about topic modelling we actually talk about genre".

Like the results of [22, 25, 24], our results give pause to consider that topics generated through topic modelling techniques are much more related to signals of genre than to semantic fields that literary researchers would consider topical and relevant. Similarly we consider that although geography can be topical for a novel, geography related signals seem much stronger than their relevance for literary analysis would warrant. Most salient is the observation that, even though we followed [25] in carefully removing function words and author specific vocabulary, we still find that topics strongly coincide with author if we recolor figure 6 according to author (see figure 7). On the one hand this may mean that authors keep within genre, on the other it means it is still hard to decide what topic model topics relay to us.

So far, we have only looked at one topic modelling technique (Top2Vec) and two segment/document sizes. Additionally, our findings are limited because we used only Dutch NUR coding as a genre target. For now, we have also disregarded the skewed makeup of our corpus, in which the romance genre is severely over-represented (cf. figure 3). We still need to evaluate the effects of using a different corpus balancing, different document sizes, isolating subgenres, different topic modelling techniques such as classic LDA, and different genre labels.

In our current corpus, topic turns out to be strongly associated with genre as labeled by

Dutch publishers. Our next step will be to determine the distribution of topics across different NUR genres. After that we aim to gauge how features of reader reviews relate to the topics we found.

References

[1] M. J. Adler. *The Great Ideas: A Syntopicon of Great Books of the Western World.* Vol. 2. Encyclopaedia Britannica, 1952.

[2] M. Algee-Hewitt, R. Heuser, and F. Moretti. *Stanford Literary Lab Pamphlet 10: On Paragraphs. Scale, Themes, and Narrative Form.* 2015. URL: https://litlab.stanford.edu/LiteraryLabPamphlet10.pdf.

[3] D. Allington. "Customer Reviews of 'Highbrow' Literature: A Comparative Reception Study of The Inheritance of Loss and The White Tiger". In: *American Journal of Cultural Sociology* 9.2 (2021), pp. 242–268.

[4] D. Angelov. "Top2Vec: Distributed Representations of Topics". In: *arXiv preprint arXiv:2008.09470* (2020).

[5] C. Baldrick. *The Oxford Dictionary of Literary Terms [online].* Oxford University Press, 2015.

[6] K. Bode. ""Man people woman life" - "Creek sheep cattle horses": Influence, Distinction, and Literary Traditions". In: *A World of Fiction: Digital Collections and the Future of Literary History.* University of Michigan Press, 2019, pp. 157–197.

[7] R. S. Buurma. "The Fictionality of Topic Modeling: Machine Reading Anthony Trollope's Barsetshire Series". In: *Big Data & Society* 2.2 (2015), p. 2053951715610591.

[8] R. Egger and J. Yu. "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts." In: *Frontiers in sociology* 7 (2022), p. 886498. DOI: 10.3389/fsoc.2022.886498.

[9] A. Goldstone and T. Underwood. "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us". In: *New Literary History* 45.3 (2014), pp. 359–384.

[10] A. Goldstone and T. Underwood. *What Can Topic Models of PMLA Teach Us About the History of Literary Scholarship?* 2012. URL: https://tedunderwood.com/2012/12/14/what-can-topic-models-of-pmla-teach-us-about-the-history-of-literary-scholarship/.

[11] R. Heuser and L. Le-Khac. *A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method, Literary Lab Pamphlet 4.* 2018.

[12] K. Jautze, A. van Cranenburgh, C. Koolen, et al. "Topic Modeling Literary Quality". In: *Dh.* 2016, pp. 233–237.

[13] M. L. Jockers. *Macroanalysis: Digital Methods and Literary History.* University of Illinois Press, 2013.

[14] M. L. Jockers and D. Mimno. "Significant Themes in 19th-Century Literature". In: *Poetics* 41.6 (2013), pp. 750–769.

[15] M. Lundy. "Text Mining Contemporary Popular Fiction: Natural Language Processing-Derived Themes Across Over 1,000 New York Times Bestsellers and Genre Fiction Novels". PhD thesis. University of South Carolina, 2020.

[16] L. McInnes, J. Healy, and S. Astels. "HDBscan: Hierarchical Density Based Clustering". In: *Journal of Open Source Software* 2.11 (2017), p. 205.

[17] L. McInnes, J. Healy, N. Saul, and L. Großberger. "UMAP: Uniform Manifold Approximation and Projection". In: *Journal of Open Source Software* 3.29 (2018), p. 861.

[18] D. McIntyre and D. Archer. "A Corpus-based Approach to Mind Style". In: (2010).

[19] J. Misset. *"Replete with instruction and rational amusement"?: Unexpected Features in the Register of British Didactic Novels, 1778–1814.* 2022.

[20] F. Pianzola, S. Rebora, and G. Lauer. "Wattpad as a Resource for Literary Studies. Quantitative and Qualitative Examples of the Importance of Digital Social Reading and Readers' Comments in the Margins". In: *PloS one* 15.1 (2020), e0226708.

[21] E. Saral and R. G. Alhama. "A Topic Modeling Study of the COVID-19 Impact in an Online Eating Disorder Community in Reddit". In: Tilburg: Tilburg University, 2022. URL: https://clin2022.uvt.nl/a-topic%20-modeling-study-of-the-covid-19-impact-in-an-online-eating-disorder-community-in-reddit/.

[22] C. Schöch. "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama." In: *DHQ: Digital Humanities Quarterly* 11.2 (2017).

[23] J. Schröter and K. Du. "Validating Topic Modeling as a Method of Analyzing Sujet and Theme." In: *Journal of Computational Literary Studies* 1 (2022).

[24] L. Thompson and D. Mimno. "Authorless Topic Models: Biasing Models Away from Known Structure". In: *Proceedings of the 27th International Conference on Computational Linguistics.* Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018, pp. 3903–3914. URL: https://aclanthology.org/C18-1329.

[25] I. Uglanova and E. Gius. "The Order of Things. A Study on Topic Modelling of Literary Texts." In: *Chr* 18-20 (2020), p. 2020.

[26] K. Van Rees, S. Janssen, and M. Verboord. "Classificatie in het culturele en literaire veld 1975-2000: Diversificatie en nivellering van grenzen tussen culturele genres". In: *Productie van literatuur. Het literaire veld in Nederland 1800-2000.* Ed. by G. Dorleijn and K. Van Rees. Nijmegen: Vantilt, 2006, pp. 239–283.

[27] "Nur". In: *Algemeen letterkundig lexicon.* Ed. by G. Vis, P. Verkruijss, H. Van Gorp, D. Delabastita, G. Van Bork, L. Bernaerts, F. Willaert, E. Op de Beek, and N. Geerdink. Digitale Bibliotheek voor de Nederlandse Letteren, 2012. URL: https://www.dbnl.org/tekst/dela012alge01%5C%5F01/dela012alge01%5C%5F01%5C%5F01441.php.

[28] M. Walsh and M. Antoniak. "The Goodreads "Classics": A Computational Study of Readers, Amazon, and Crowdsourced Amateur Criticism". In: *Journal of Cultural Analytics* 4 (2021), pp. 243–287.

[29] M. Wilkens. “Genre, Computation, and the Varieties of Twentieth-Century U.S. Fiction”. In: *CA Journal of Cultural Analytics* 2.2 (2016). DOI: 10.22148/16.009. URL: https://culturalanalytics.org/article/11065.

A. Unit of Measure

A.1. Segmentation

From the perspective of literary studies, it is illogical to bind topic to the full text of a novel. A novel likely touches on a multitude of topics, so a division into chapters, sections, paragraphs or even sentences might yield more useful topics. However, segmenting whole novels into minimum-sized windows has consequences for the co-occurrence of words and the number of topics that will be detected. By segmenting a whole novel, the words within one segment no longer co-occur with the words in another segment from the same novel. Therefore, the word co-occurrence matrix becomes more sparse, so topic modelling algorithms identify more and smaller dense clusters, resulting in more topics.

We investigate the impact of segmenting on the co-occurrence of words by analysing how the number of co-occurring pairs of lemmas increases as we iterate over novels, using either the whole novels as document boundaries, or segments constructed from joining a sequence of paragraphs in a novel into segments containing at least 5,000 words. Figure 8 shows the result. The X-axis shows the number of lemmas (after filtering out the most and least frequent lemmas as described above) and the Y-axis shows how many distinct co-occurring pairs of lemmas are found.

The difference between segmenting and not segmenting results generates a clear picture. At the level of whole novels, well over 196 million co-occurrence pairs are indexed after having seen 300,000 lemmas (corresponding to about 250 novels). For the segmented novels, at the

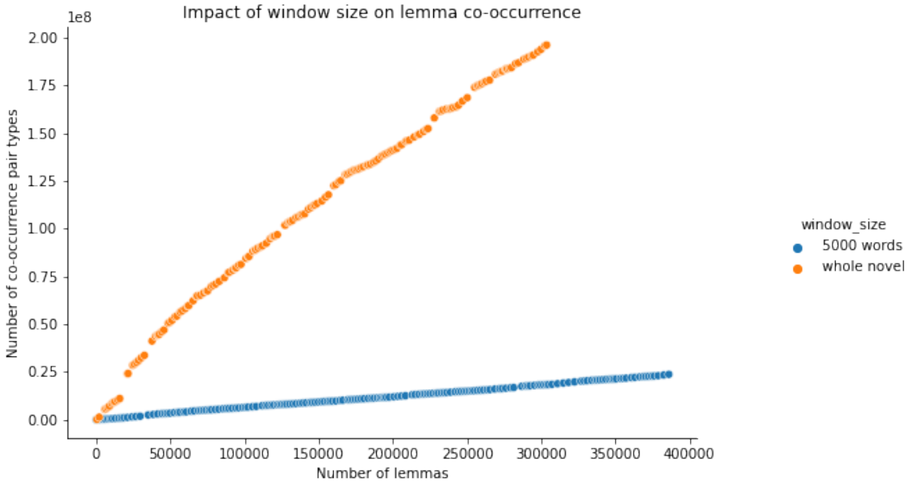

Figure 8: The difference between full novel and a window size of 5,000 tokens on the number of co-occurring lemmas

same 300,000 lemma point, there are only 18.3 million pairs. That is a full order or magnitude less. Regarding a choice for unit of measure, this means that we have to carefully investigate the trade-off that exists between the size of documents fed to a topic modelling algorithm (i.e. whole novels or smaller segments) and the number of topics returned.

B. Topic examples

The following are examples from topics, generated by Top2Vec at the level of whole novels, that show a high concentration of coherent geographical lemmas. Lemmas strongly related to a coherent geographical location are in bold.

Topic number: 4

nochtans, komaan, stilaan, gsm, parking, job, vooraleer, miserie, **kot**, **brussels**, **antwerps**, euh, proper, gsmnummer, **goesting**, bijgevolg, verdict, plezant, **antwerpen**, **vlaams**, voormiddag, speurder, voordien, **zaventem**, **oostende**, flik, ontgoocheling, zogezegde, deurgat, **gent**, autosnelweg, voorhand, pv, **gents**, crapuul, evolueren, zijt, **knokke**, **vlaming**, sukkelaar, **mechelen**, recupereren, meermaals, contacteren, evident, nonkel, allez, klasseren, rijkswacht, **schelde**

Topic number: 5

vondelpark, **schiphol**, **leidseplein**, bitterbal, shag, know, like, lullen, please, gadverdamme, it, **amsterdams**, snot, never, **grachtenpand**, can, **kroket**, **amsterdam**, goor, spuug, there, veegt, kut, ie, see, bh, **amstelveen**, fietspad, plee, geilheid, wc, gezeik, rouwkaart, quote, **almere**, sure, tje, fucking, is, ehm, **hilversum**, only, **randstad**, **drop**, lacherig, **koninginnedag**, too, zeiken, opschuden, poep

Topic number: 11

zo, polder, hbs, jenever, **haarlem**, zoldering, stationsplein, ballpoint, tramhalte, **schevenings**, verveloos, **scheveningen**, **vondelpark**, shag, **rotterdam**, waartussen, grammofoon, **zandvoort**, **rijksdaalder**, **arnhem**, hongerwinter, jongensboek, **ijsselmeer**, gymnasium, schemer, schoolschrift, ijl, schemren, brokkelig, stofjas, **amstel**, klomup, vitrage, bakeliet, sigarenwinkel, schrijfmachine, vergelen, bovenhuis, rui, plantsoen, brilleglas, **groningen**, celluloid, windstil, trapper, **leidseplein**, vooroorlogs, veraf, allengs, **wassenaar**

Topic number: 15

stockholm, **zweden**, **kronen**, **kopenhagen**, **oslo**, **noorwegen**, **denemarken**, **deens**, **noors**, midzomer, **fins**, **zweed**, vooronderzoek, line, verhoren, ordner, politieacademie, lichtkegel, legitimatie, volvo, **finland**, strafregister, rechercheteam, **fjord**, kelderruimte, nor, wide, politiemen, hoofdbureau, moordzaken, geweldsdelict, villawijk, **scandinavisch**, messing, goddomme, avondkrant, rijkweg, smeriss, opsporing, bewakingscamera, **scandinavie**, freule, pagekapsel, nova, zomerhuis, **oostzee**, sankt, thomas, moordonderzoek, sterfgeval