Faces, Fights, and Families: Topic Modeling and Gendered Themes in Two Corpora of Swedish Prose Fiction

Mats Dahllöf[0000−0002−4990−7880] and Karl Berglund[0000−0001−7280−1112]

Uppsala University, Uppsala, Sweden
mats.dahllof@lingfil.uu.se, karl.berglund@ub.uu.se

Abstract. This paper explores topic modeling (TM) as a tool for "distant reading" of two Swedish literary corpora. We investigate what kinds of insight and knowledge a TM-based approach can provide to Swedish literary history, and which methodological difficulties are associated with this endeavour. The TM is based on 12- and 24-term chunks of selected verb and common noun lemmas. We generate models with 20, 40, and 100 topics. We also propose a method for a quantitative and qualitative *gendered thematic analysis* by combining TM with a study of how the topics relate to gender in characters and authors. The two corpora contain, respectively, Swedish classics (1821–1941) and recent bestsellers (2004–2017). We find that most of the topics proposed by the TM are easy to interpret as conceptual themes, and that the "same" themes appear for the two corpora and for different TM settings. The study allows us to make interesting observations concerning different aspects of gender and topic distribution.

Keywords: Topic Modeling · Distant Reading · Gender Analysis · Literary Methodology · Swedish Prose Fiction · Bestsellers.

1 Introduction

The aim of this paper is to explore topic modeling (TM) as a tool for "distant reading" of Swedish literary corpora. We want to investigate what new kinds of insight and knowledge a TM-based approach can help us gain as regards Swedish literary history. We also want to discuss some of the methodological difficulties associated with this endeavour. In particular, this article proposes an approach to quantitative and qualitative *gendered thematic analysis* by combining TM with a study of how the topics relate to gender in characters and authors. This has, as far as we know, not been tried before.

Our aims are exploratory and mostly focused on methodical investigation, but results will also be reported and discussed. The study is concerned with two corpora: prose fiction with modern(ized) Swedish spelling from Litteraturbanken (mainly Swedish classics, 1821–1941); and prose fiction from contemporary Swedish bestseller charts (2004–2017).

Our research questions can be stated as follows:

- How well does TM work as a tool for extracting content themes from Swedish literary corpora?
- How robust and reproducible are the results of a TM system?
- Is it possible to find connections between topics and the gender of characters and authors?
- What are the advantages of using TM as a tool for the analysis of (Swedish) literature in comparison with other methods?

2 Literature and Topic Modeling – State of the Art

In traditional studies on themes in literature, the researcher approaches the texts having some predefined theme as his or her point of departure. This choice might be motivated by e.g. its historical, stylistic/aesthetic, or political significance. The established methodology in finding themes to investigate is to rely on already read books, on books that could or should be relevant, or on previous research. In recent times, free-text search engines have come to be more and more used for locating instances of themes in literary corpora. Still, such procedures rely on the researcher's assumptions about which terms are indicative of the relevant themes.

As several scholars of literature have already pointed out, topic modeling (TM) makes it possible to approach thematic literary analysis from another angle. Instead of first deciding on a theme and a material and then search for passages expressing the theme, the researcher selects a collection of texts and makes use of an algorithm to find which "topics" – as components of a latent probabilistic model – best explain the structure of the texts, e.g. [10, 33]. Such a bottom-up approach to thematic literary analysis will generate proposals which are quantitatively justified by the data in a way that is not possible in the traditional frameworks of literary studies.

The present work belongs to a paradigm of computational quantitative large-scale analysis of literature, which Franco Moretti [22] wittily has characterized as *distant reading*. There are a few exploratory examples of this kind of literary criticism on Swedish literary material, e.g. [12, 6, 4], but only one uses TM [4]. The shift from manual qualitative narrow-scale methods to computational distant reading has been criticized by researchers in the humanities for being reductive, positivist, white male-centred, and not critical enough, e.g. [1, 15, 20, 18]. Objections have also been raised in a Swedish context, e.g. by [5, 16], despite the fact that there are only few examples of this kind of research on Swedish material.

TM and similar algorithmic methods thus appear to be both productive and provocative to literary scholars. Although we believe in the usefulness of distant reading approaches, we will discuss our results with an awareness of the methodological problems.

2.1 Generative Model

A topic model (TM, but notice the difference between model and modeling here) is a model of a collection of text pieces, which can be segmented according to different criteria. In our approach, the text segments (called *chunks*) are (typically) smaller than paragraphs and only contain a selection of lexical terms. A TM of the kind used here is a probabilistic model of how the text chunks are generated by a hypothetical stochastic process. The idea is that we can view each chunk as a sequence of terms, which is generated in such a way that we first draw a topic according to the topic distribution of the chunk, and then each term in the chunk given the term probabilities of the topic [30].

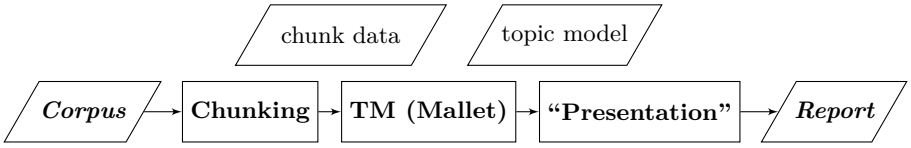


Fig. 1. The pipeline with the steps involved in our – or a typical – application of TM.

The *modeling* process consists in finding a model that as well as possible fits the data, i.e. the chunks, which reflect the surface facts of the corpus. The last step, in an approach like the present one, is to turn the statistical model into something which can be read and interpreted. We follow common practice in generating lists of the most significant words associated with each topic. The pipeline is shown in Figure 1.

In applications of TM, we typically want the topics to capture content themes. Of course, the concept of theme is a vague and open one. Themes can be more or less specific, as well as partially overlapping. Saying which theme a topic represents (based on the corresponding list of significant words) is consequently a matter of qualitative analysis. The noun *topic* is often used as a synonym to theme in everyday language, but we will only use *topic* in the technical sense of TM here. So, topics correspond to themes, if the TM is successful, but they can also fail to do so. Furthermore, topics can not be expected to be conceptually "pure", as minor subthemes can be associated with topics even if a main thematic label is clearly justified.

2.2 TM and Literature Research

Several literary scholars have recognized the relevance of TM for "distant reading". As can be expected, this research has often turned to English literature, such as poetry [27], 19th century prose fiction [10, 11], or contemporary American bestsellers [3]. Jockers [10] makes use of TM to perform narratological analysis. Among other influential entries we find work on the French *Encyclopédie*

(18th century) [28], French classical drama (17th–18th century) [29], Spanish Golden-Age sonnets [23], Danish 19th century literature [32], and a meta-study of American research on literature [9].

The only previous study using TM on Swedish literature is Barakat's [4] master thesis (in statistics and data mining). He tries to map statistically what makes audiobooks popular. Although the study is based on a large corpus (3077 books) and poses interesting questions, its focus is on statistical matters rather than literary ones. There are also a few studies using TM on non-literary Swedish material, such as governmental official reports [25, 26] and Swedish parliamentary debates and the discourse on immigration [17].

3 Method and Data

The concept of topic is a statistical one. A topic captures a "probability distribution over terms", intended to model what can be understood as "recurring themes" in a collection of text chunks [8]. Each chunk is associated with a probability distribution over the topics. TM is a process of estimating the two distributions. A TM system is guided by a number of parameters and the user's decisions on data and parameter settings are crucial for the result it will produce.

Our approach involves three steps. First, we propose a procedure, of our own invention, for selecting the terms and capturing the chunks to be fed to the TM module. Secondly, there is the TM per se, which relies on a TM module of the Mallet [21] software package. Finally, the results of the TM are presented to the user for qualitative interpretation. The discussion is based on our own reading of the results.

3.1 Gender-related Labeling

The paragraphs from which the text chunks are generated are labeled by two kinds of gender-related information. The first kind indicates whether a paragraph is only about female or only about male characters. We also label paragraphs according to the gender of the author(s). When there are more than one author, and they belong to different sexes, we do not use either label.

We operationalize the character gender distinction by labeling paragraphs containing at least one singular third person feminine pronoun (*hon* [subject form], *henne* [object form], *hennes* [genitive]), without containing a singular third person masculine pronoun (*han*, *honom*, *hans*). This gives us a simple way of identifying passages which are about female characters. We also label the paragraphs involving a singular third person masculine pronoun (but not any feminine one) in the corresponding way. The two categories are thus by definition mutually exclusive. However, many paragraphs will not belong to either of the two categories, by not containing any such pronoun or by involving both genders. Also, note that pronouns are not among the lexical terms preserved in the chunking (see next section), so they will not be "seen" by the TM.

3.2 Chunking and Selection of Lexical Terms

The first step in the processing of the texts is part-of-speech tagging.[1] Lexical terms are then formed by combining the base form and the part-of-speech tag. This means that all inflected forms are grouped as one term (lemma) and that the part-of-speech tag disambiguates some lemmas, e.g. *röra*, giving us *röraNN* (jumble) or *röraVB* (touch). Only a subset of the lexical terms is used for the TM. First, we remove all instances of the 100 most frequent ones. (They form a list of stop words, as it were.) Furthermore, we require, for the Classics corpus, that the terms have at least 10 instances distributed over at least 10 different books. The corresponding requirement for the Bestsellers corpus is 40 instances over 20 different books, which is proportional to the relevant sizes. Finally, we only include nouns and verbs among the terms, assuming that these carry a high "semantic weight", and are the most useful indicators of recurring themes.

```
supaVB veckaNN skötaVB tjänstNN församlingNN måsteVB klagaVB
prostNN biskopNN biskopNN sockenNN hållaVB

tjänstNN församlingNN måsteVB klagaVB prostNN biskopNN biskopNN
sockenNN hållaVB korNN bröstNN prästNN

visshetNN fiendeNN kyrkaNN fiendeNN bänkNN bondeNN kyrkaNN korNN
fiendeNN fiendeNN fiendeNN trampaVB
```

Table 1. The three first 12-term chunks from Lagerlöf's *Gösta Berlings saga*.

After the tagging and the compilation of a frequency dictionary for the corpus, each text is processed paragraph by paragraph. These are converted to lists of lexical terms, $t_1, t_2, \ldots, t_p$ (when the paragraph yields p terms). From these we generate the chunks which will form the data fed to the TM module. The chunks are of a predefined length, $c \in \{12, 24\}$, in our experiments. In comparison with e.g. [10], our chunks are very small, intended to focus on themes which are prominent at that level of textual "resolution". The "window" of the "chunker" moves forward three terms[2] for each capture. So, $t_1, t_2, \ldots, t_c$ will be the first chunk, $t_4, t_5, \ldots, t_{c+3}$, the second one, $t_7, t_5, \ldots, t_{c+6}$, the third one, and so on, until $t_{3n+1}, t_{3n+2}, \ldots, t_{3n+c}$, for the largest n such that $3n + c \leq p$. (So, this procedure gives us $n + 1$ chunks for a paragraph term sequence of length p.) See Table 1 for an example. We do not allow chunks to extend over paragraphs, which we assume are often associated with shifts in thematic content. We ignore sentence boundaries. Also note that the chunks overlap and that the number of term instances in the chunk set will be considerably larger than in the actual text.

[1] We use Stagger [34] A few obvious frequent tagging errors are corrected.
[2] Three rather than one for "economic" reasons.

3.3 Topic modeling

The TM is done by means of the *ParallelTopicModel* class of the Mallet [21] software. The class is characterized (in a code comment) as providing a "[s]imple parallel threaded implementation of LDA [Latent Dirichlet Allocation], following [24], with SparseLDA sampling scheme and data structure from [13]". The output is strongly influenced by the setting of a number of parameters. We generated a fairly small number of topics, 20, 40, or 100. Our settings are consequently geared towards the extraction of general themes (see [10]). The Mallet settings were based on tuning on the Classics corpus and manual inspection of the results.

$k \in \{20, 40, 100\}$ [a.k.a *numTopics*]	$numIterations = 2000$
$alphaSum = 2.0 * k$	$burninPeriod = 200$
$beta = 0.001$	$optimizeInterval = 100$
$symmetricAlpha = false$	

The high $alphaSum = 2.0 * k$ is motivated by a wish to avoid "favoring just a few topics" [30], whereas *beta*, which "smoothes the word distribution in every topic" [30] is given a low value (0.001). Since the set-ups stipulate that number of topics is fairly small, these will be of a general nature. The chunking will furthermore promote the extraction of themes that manifest themselves locally in texts. The small *beta* will promote models where terms tend to be specific to a few topics.

3.4 Presentation of the TM Output

The result of the TM is essentially an assignment of a topic (identified by an integer) to each instance of a term in the chunks. This means that we can estimate, for each term (type) and topic, the probability that the term represents the topic, p(topic|term), for the complete corpus. Similarly, the result defines, for each chunk and topic, a ratio saying to what degree the chunk represents the topic, p(topic|chunk). As the chunks derive from literary works, and are labelled with gender-related information, we can also compute how large a share a topic has in a particular book, and in male and female authors, and in passages with male and female characters.

As is common practice, we present the topics for "reading" as lists of "top" terms. We rank the terms according to the χ^2 (chi square) statistics,[3] which quantifies the strength of association between a term and a topic [19]. We think that this is better than looking at the more "elementary" conditional probabilities: p(term|topic) is high for frequent, but consequently general words. And p(topic|term) will be close to 100% for many very rare – and consequently

[3] It is computed in this way: $\chi^2 = \sum_{F \in \{f, \bar{f}\}} \sum_{T \in \{t, \bar{t}\}} \frac{(N_{FT} - E_{FT})^2}{E_{FT}}$, where N_{FT} is the actual number of observations and E_{FT} the expected number of instances under the assumption that T and F are independent. Here, χ^2 is used to generate word lists from the TM output. (We do not use it for significance testing.)

not very significant – words. We do however use colour and boldface to indicate p(topic|term), since this probability tells us how specific a term is for a topic/theme. The styles are to be read as follows: **TermPOS** $\geq 90\%$, otherwise **TermPOS** $\geq 75\%$, otherwise **TermPOS** $\geq 50\%$, otherwise (plainly) TermPOS ($< 50\%$), see e.g. Table 3.

3.5 Data: the "Classics" and the "Bestsellers" Corpora

We have applied our method for TM and gender analysis on two quite different corpora of Swedish prose fiction, "Classics" and "Bestsellers". The "Classics" corpus is curated from Litteraturbanken (LB) (www.litteraturbanken.se), which is a collection of Swedish literature mainly from the 19th century and the first decades of the 20th century. The focus of LB is on classics and literature of particular historical or aesthetical importance. It also contains translations and reference literature [14]. LB comprises more than 700 e-texts, as well as facsimile editions. The Classics corpus is a subset of the e-texts, which covers the prose fiction with modern(ized) (post-1906) spelling, as spelling variation would interfere with the TM. With duplicate editions removed, the corpus consists of 121 Swedish novels or collections of short stories (6.6 million words), originally published between 1821–1941, the majority after 1900. Male writers are over-represented: Only 36% of the works are by female authors. (The full list of works is available as an appendix to the TM presentations among the Supplementary Materials.)

Since LB is centred around some influential Swedish authors and their works, the corpus is not representative of the full range of literary production in Sweden during the period in question. Authors like Hjalmar Bergman (14 works), Selma Lagerlöf (16), and August Strindberg (23) are over-represented, while others are not found at all. Nevertheless, the LB corpus allows us to study the recurring thematic content in some of the most prominent Swedish writers of the time. The Classics corpus was compiled from the system-internal XML-files of LB.

The second corpus studied in this article, "Bestsellers", is based on the Swedish bestseller charts compiled by the book trade magazine *Svensk Bokhandel* [31]. We collected all prose fiction on the lists 2004–2017 in the two categories "bestselling hardbound fiction" and "bestselling paperbacks". We then excluded all non-fiction, all fiction that does not contain prose (just a few works of poetry), and all duplicates. (Duplicate entries are common since bestsellers tend to sell well both in hardbound and paperback editions.) This gives us 280 bestselling novels and collections of short stories, of which 231 works ($82,5\%$) were available in digital form. The raw text from these, some 25.7 million words, thus constitutes the corpus.[4] This corpus covers more than four fifths of all bestselling prose fiction published in Swedish during the first decades of the 21st century.

[4] The extraction and processing of the text were conducted within the confines of the activities of the university library. Everything except the literary text proper has been removed, including meta-data, dedications, forewords, afterwords, acknowledgements, extra material, etc.

The 49 titles that were not available in digital form seem to be quite evenly distributed over time and genres. However, Swedish crime fiction is presumably somewhat over-represented in the corpus and many titles are missing from 2017.

The bestseller corpus is dominated by novels (there is only one collection of short stories), Swedish originals (75%), and crime fiction (62%). However, it also includes translated fiction, genre fiction other than crime (romance, science fiction, fantasy), as well as literary and more aesthetically experimental fiction. In contrast to the Classics corpus, the gender distribution in the Bestseller corpus is balanced: 48% of the works are written by female authors.

	Classics Corpus						Bestsellers Corpus					
Author	WC	FP	MP	Chunks	RF	RM	WC	FP	MP	Chunks	RF	RM
Female	2.4	25.3	26.8	51508	12082	14917	11.1	25.5	24.0	233733	62755	59886
Male	4.2	10.2	29.6	148517	13757	74045	13.5	15.7	27.4	253228	40166	110616
"Both"	-	-	-	-	-	-	1.1	26.2	25.9	10625	2496	3613
Sum	6.6	-	-	200025	25839	88962	25.6	-	-	497586	105417	174115

Table 2. Some corpus statistics: WC: word count in millions. FP: relative frequency of singular third person feminine pronouns, in per mille. MP: ditto masculine. Chunks: size of the set of 12-word chunks. RF: size of female character subset. RM: ditto male. (See section 3.1.)

If we look at the sizes of the data converted into 12-word chunks, we get the numbers in Table 2. We immediately see that the female authors write about male and female characters in a balanced way, whereas the male authors write about male individuals three to four times more often than about women.

Considering their sizes and principled composition, the two corpora provide a solid point of departure for a study on thematic trends in Swedish literature (including recent translated bestsellers). They also allow us to compare, from a thematic point of view, literature from the beginning of the 20th century with literature from the beginning of the 21st century.

3.6 Experiments

The result of a TM experiment is crucially dependent on the user's design choices, such as chunk size, number of topics, and assumptions about distribution priors. Our design and parameter tuning were based on experiments with the Classics corpus, 12- and 24-term chunks, and $k \in \{20, 40\}$. We then applied the settings that seemed useful to the Bestsellers corpus and to experiments with $k = 100$. At URL `https://stp.lingfil.uu.se/~matsd/dhn2019`, we provide as Supplementary Materials for this artice, the automatically generated reports from our TM experiments, with listings of the works included in the corpora.

4 Results and Discussion

We would like to claim that the topics proposed by our experiments to a high degree are possible to interpret in a "natural" way as expressions of one main general theme. However, many topics seem to contain traces of subordinate themes. Consider, for instance, topic #7 from the Bestsellers corpus when $k = 20$ and $c = 24$, as exhibited in Table 3. It seems (also when we look at the longer wordlist) that fighting and violence is the central theme. But we also find nouns for categories of animal. This can be explained by the narratives in the corpus, but the topic is probably also associated with chunks featuring peaceful animals and fights without animals. Because of this thematic polysemy and vagueness, it is often not obvious which terms to use in labeling the topics. We have mixed two "strategies": In some cases we use a general term that covers the central part of the topic. For other topics, we prefer a conjunction of more specific terms. (The reader is invited to look at the Supplementary Material to get a fuller picture.)

In the experimental set-ups with $k = 20$, for both corpora, 15 intuitively identified themes on different levels of specificity appeared as topics in all cases. So at $k = 20$, three out of four themes were the "same" for $c \in \{12, 24\}$. We also tried our scheme on the 2010–2017 subset of the Bestsellers corpus: With $k = 20$ and $c = 24$, again roughly three out of four topics seemed to capture the "same" ones as those extracted from the full Bestsellers corpus. Our experiments also allowed us to find remarkable thematic similarities between the two corpora, see Section 4.2. These results suggest that the current TM approach has a certain degree of stability.

Changes in the number of topics obviously lead to changes in the output. In many cases we see that more general themes in a conceptually reasonable way split into more specific ones when we increase the number of topics from $k = 20$ to $k = 40$, and then further at $k = 100$. For instance, a persistent theme surfacing for the Bestsellers corpus is one corresponding to action-loaded fighting scenes. With $k = 20$ this theme is captured by one topic (see Table 3), which seem to derive from many kinds of fighting scene, mostly from crime fiction, but also from other genres (e.g. historical fiction, which is a plausible explanation for the high ranking of *horse*). The theme is quite clearly split in two at $k = 40$, one related to blunt force and another to firearm violence (see Table 3). And when we set $k = 100$, this general theme yields four different topics (see Table 3). Here, topic #99 is clearly related to firearms ($c = 24$, henceforth). Topic #55 seems to relate to more serious violence, involving e.g. knives or subsequent injuries. Topics #19 and #78 both seem to capture fist fights, but at slightly different stages: Topic #19 is related to the middle stages of fights, whereas topic #78 corresponds to their consequences.

These results show that the TM has the ability to discover similarities between different kinds of fight when the number of topics is small, but that it is also able to separate different kinds or aspects of fights from one another as the number of topics grows larger. There are other examples. Topics inviting the labels "Outdoor settings" and "Eating and drinking" from the Bestsellers corpus behave in a similar fashion: Content related to outdoor settings is gath-

$k = 20$

#7 **skjutaVB**, **hästNN**, **skrikaVB**, **knivNN**, **markNN**, **vapenNN**, **skottNN**, fallaVB, springaVB, **kulaNN**, **djurNN**, **pistolNN**, **sparkaVB**, **pilNN**, slåVB, hundNN, fotNN, **gevärNN**, **slitaVB**, **skrikNN**. [*shoot, horse, shout, knife, ground, weapon, shot, fall, run, bullet, animal, pistol, kick, arrow, hit, dog, foot, gun, tear, cry.*]

$k = 40$

#8 **skrikaVB**, slåVB, **sparkaVB**, tagNN, fallaVB, **slitaVB**, armNN, **tappaVB**, kastaVB, huvudNN, fotNN, benNN, **balansNN**, **gripaVB**, **snubblaVB**, **näveNN**, **sparkNN**, **vrålaVB**, kraftNN, **skrikNN**. [*shout, hit, kick, hold, fall, tear, arm, drop, throw, head, foot, leg, poise, seize, stumble, fist, kick, roar, force, shout.*]

#22 **skjutaVB**, **springaVB**, **stegNN**, **sekundNN**, **skottNN**, **vapenNN**, **pistolNN**, **kulaNN**, vändaVB, **revolverNN**, stannaVB, **gevärNN**, fotNN, meterNN, ögonblickNN, **siktaVB**, rörelseNN, närmaVB, hinnaVB, rusaVB. [*shoot, run, step, second, shot, weapon, pistol, bullet, turn, revolver, stop, rifle, foot, meter, moment, aim, movement, approach, find time, rush.*]

$k = 100$

#19 **fallaVB**. **kastaVB**, slåVB, golvNN, **markNN**, **ramlaVB**, knäNN, **landaVB**, **kantNN**, **krossaVB**, brytaVB, **vältaVB**, hoppaVB, rullaVB, stötaVB, **studsaVB**, **bakhuvudNN**, vikaVB, träffaVB, **slungaVB**. [*fall, throw, hit, floor, ground, tumble, knee, land, edge, crush, break, overthrow, jump, roll, push, bounce, back of the head, fold, hit, hurl.*]

#55 **tagNN**, **smärtaNN**, **knivNN**, **gripaVB**, **halsNN**, **skäraVB**, **tandNN**, armNN, **käkeNN**, **knäckaVB**, **muskelNN**, **huggaVB**, **ormNN**, **bitaVB**, ryggNN, sparkaVB, **snaraNN**, **sprutaNN**, **ryggradNN**, greppNN. [*grip, pain, knife, grab, throat, cut, tooth, arm, jaw, break, muscle, chop, snake, bite, back, kick, noose, syringe, spine, grip.*]

#78 **fotNN**, **benNN**, **balansNN**, **tåNN**, **tappaVB**, **trampaVB**, krypaVB, släppaVB, **vacklaVB**, **greppNN**, **rockNN**, resaVB, **vristNN**, fläckNN, **spindelNN**, **högerhandNN**, **svajaVB**, **stampaVB**, **haltaVB**, **hasaVB**. [*foot, leg, balance, toe, loose, tread, crawl, drop, falter, grip, coat, travel/rise, ankle, spot, spider, right hand, swing, stomp, limp, shamble.*]

#99 **skjutaVB**, **vapenNN**, **skottNN**, **pistolNN**, **kulaNN**, **gevärNN**, **revolverNN**, **siktaVB**, **magasinNN**, **patronNN**, **kornNN**, **avlossaVB**, **ammunitionNN**, riktaVB, **avtryckareNN**, **avfyraVB**, **kaliberNN**, **laddaVB**, **containerNN**, **bågeNN**, **kolvNN**. [*shoot, weapon, shot, pistol, rifle, revolver, aim, magazine, cartridge, sight (firearm), trigger, ammunition, aim, trigger, trigger, calibre, load, container, bow, cylinder.*]

Table 3. Topics from the Bestseller corpus. A "Fighting, violence, and animals." theme splits into more specific topics as the number of topics (k) increases, in all cases $c = 24$.

ered in one topic #6 when $k = 20$ (again $c = 24$), split in three when $k = 40$, one capturing settings adjacent to water #20, one houses and gardens #28, and one woods and fields #37. It yields four topics when $k = 100$. We also find a clear "Eating and drinking" topic #14, when $k = 20$. This is split into two when $k = 40$ and into three when $k = 100$: coffee breaks #48, setting the table #83, and drinking #87.

Outcomes like these indicate that selecting a certain number of topics does not necessary make the result more or less conceptually "correct", but rather guides the TM process to capture themes on different levels of specificity. Generally, using TM with fewer topics will capture broader thematic categories (such as fights or depictions of landscapes), whereas a larger number of topics will allow the TM to capture more specific categories (such as shootings or depictions of woods).

4.1 Topics, Themes and Gender

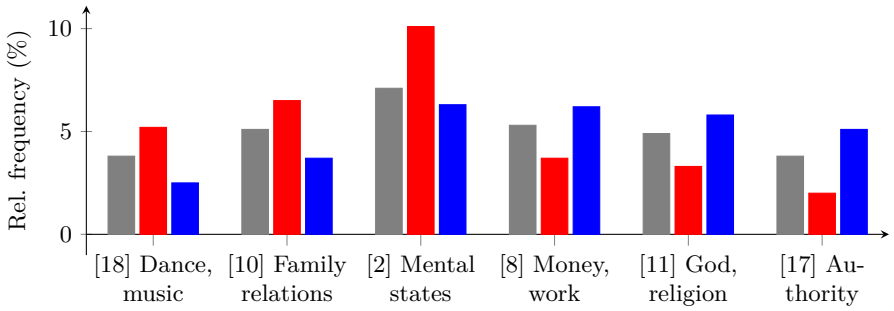
There have been previous attempts to connect themes retrieved by TM to author gender [10]. There are also large-scale studies of character gender in literary text based on pronoun evidence [7]. However, no one seems to have combined these modes of analysis. Our experimental design allows us to find gender-biased themes both as regards characters and authors.

The corpus data and the TM output make it possible, for each topic t, to compute its relative frequency in character-based female chunks $r_{t,F}$, and the corresponding value, $r_{t,M}$, for male chunks. The ratio between the two values, i.e. $r_{t,F}/r_{t,M}$, is a measure of how much more common the topic is in female compared to male contexts. We can use this number to rank the topics on a scale from highly female to highly male. (We should remember, that masculine pronouns and male character paragraphs are much more common than their female counterparts in both corpora, notably in the Classics – see Table 2.)

For the topics retrieved from the Classics corpus with $k = 20$ and $c = 24$, the most female-biased ones can be labeled "Dance, music, entertainment", "Family relations" and "Mental states, existential reflection" (see Figure 2). The three leading topics on the male end can be put under the headings "Authority" (church and nation), "God, religion, faith", and "Money, work, trade" (buying, selling, and monetary matters).

If we carry out the same exercise for the Bestsellers corpus (again with $k = 20$ and $c = 24$), we end up with the picture in Figure 3. The most female topic collects terms relating to dress, hair, and make-up, which we label "Appearance". The following two invite the labels "Family circle. Disease, health care", as there appears to be a strong connection between family members and hospitals, and "Eating and drinking", for which the top terms agree with that label in an obvious way. The topics most strongly associated with male characters concern "War, nations, global politics", "Police work, investigations, law enforcement" and "Fighting, violence. Animals" (discussed above). We also find topics with a gender-independent distribution in both corpora, for instance, "Telephone calls"

#10 in the Bestsellers corpus or "Disease, health care. Accidents" #5 in the Classics corpus.



#18 **spelaVB**, flickaNN, **dansaVB**, **sjungaVB**, **dansNN**, fruNN, **mammaNN**, **musikNN**. [*play, girl, dance, sing, dance, wife/Mrs, mother, music.*]

#10 **barnNN**, **morNN**, **farNN**, **sonNN**, **hemNN**, **förälderNN**, årNN. [*child, mother, father, son, home, parent, year.*]

#2 **tankeNN**, **drömNN**, **självNN**, livNN, **känslaNN**, **minneNN**, **verklighetNN**. [*thought, dream, soul, life, emotion, memory, reality.*]

#8 **pengNN**, **köpaVB**, **betalaVB**, **säljaVB**, **arbeteNN**, **skaffaVB**, **kronaNN**. [*money, buy, pay, sell, work, get, crown.*]

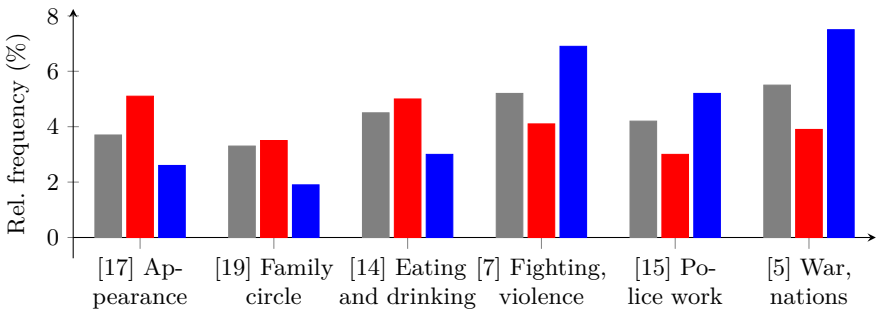
#11 **gudNN**, människaNN, **gärningNN**, **helveteNN**, **andeNN**, **djävulNN**, **nådNN**. [*God, man, deed, hell, spirit, devil, grace.*]

#17 **kyrkaNN**, **prästNN**, **konungNN**, **kungNN**, **församlingNN**, folkNN, **svenskNN**. [*church, priest, king, parish, people, Swede.*]

Fig. 2. Most gender-biased topics in the Classics corpus ($k = 20$ and $c = 24$). Red staples: $r_{t,F}$. Blue staples: $r_{t,M}$. Grey staples: relative frequency of the topic in the full corpus. Topics numbered according to their relative frequency. The top seven term of the topics are listed.

The patterns we find seem to reflect traditional gender roles and stereotypes to a high degree: To deal with authority, God, and trade was primarily responsibilities for men around the turn of the century 1900, whereas entertainment, the family circle, and reflection on mental states were more closely connected to female concerns. In contemporary bestsellers, men dominate (or are expected to dominate) in "arenas" like crime, in particular violent crime, the police force, politics, and the military. Women, on the other hand, are to a higher degree depicted in contexts where appearance, family life, and meals play a prominent role. Hence, gender roles have changed – yet they seem to remain the same in many respects.

In general, the gender bias of characters and authors point in the same direction. This is true for the examples we have given above. As can be expected, men



#17 **hårNN**, **klänningNN**, **skjortaNN**, **skoNN**, **kjolNN**, **byxaNN**, **jeansNN**. [*hair, dress, shirt, shoe, skirt, trousers, jeans.*]

#19 **pappaNN**, **mammaNN**, **svaraVB**, **frågaVB**, **mormorNN**, **sjuksköterskaNN**, **morfarNN**. [*dad, mum, answer, question, maternal grandmother, nurse, maternal grandfather.*]

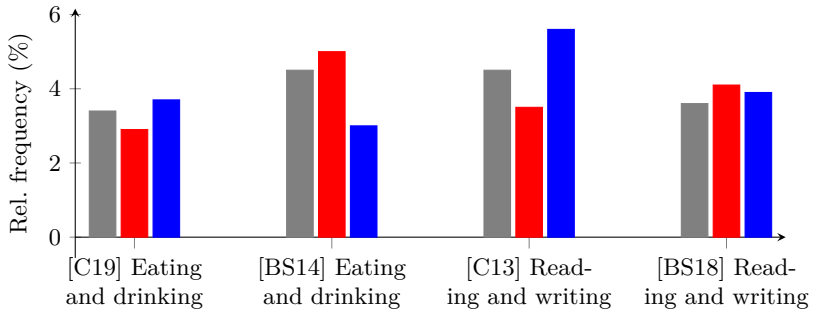
#14 **kaffeNN**, **dricka**VB, ätaVB, **matNN**, **vinNN**, **koppNN**, **glasNN**. [*coffee, drink, eat, food, wine, cup, glass.*]

#7 **skjutaVB**, **hästNN**, **skrikaVB**, **knivNN**, **markNN**, **vapenNN**, **skottNN**. [*shoot, horse, shout, knife, ground, weapon, shot.*]

#15 **polisNN**, **mordNN**, **utredningNN**, **mördareNN**, **offerNN**, **brottNN**, **vittneNN**. [*police, murder, investigation, murderer, victim, crime, witness.*]

#5 **krigNN**, landNN, **mrNN**, **presidentNN**, **distriktNN**, **arbetsgivareNN**, **nådNN**. [*war, country, mr, president, district, employer, grace.*]

Fig. 3. Most gender-biased topics in the Bestseller corpus. Read as Figure 2.



#C19 **drickaVB**, **kaffeNN**, **matNN**, ätaVB, **vinNN**, **middagNN**, **glasNN**. [*drink, coffee, food, eat, wine, dinner, glass.*]

#BS14 **kaffeNN**, **drickaVB**, ätaVB, **matNN**, **vinNN**, **koppNN**, **glasNN**. [*coffee, drink, eat, food, wine, cup, glass.*]

#C13 **läsaVB**, **bokNN**, **skrivaVB**, **tidningNN**, **professorNN**, **författareNN**, **brevNN**. [*read, book, write, newspaper, professor, author, letter.*]

#BS18 **skrivaVB**, **bokNN**, **läsaVB**, **brevNN**, **bildNN**, **textNN**, **papperNN**. [*write, book, read, letter, picture, text, paper.*]

Fig. 4. Themes were gender biased has changed. **C** for Classics and **BS** for Bestsellers; otherwise as Figure 2.

generally tend to write more about male-biased themes and women more about female-biased themes. This also agrees what is known from previous research. There are however a few exceptions to this pattern. The theme "Agriculture, animals" #16 in the Classics corpus and "Money and trade" #20 in the Bestsellers corpus are clearly male-biased in the texts, but have higher or equal representation among female authors. The opposite pattern can be noted in "Dance, music, entertainment" #18 in the Classics corpus and "Family circle. Disease, health care" #19 in the Bestsellers corpus, where the topics are more prevalent in female character paragraphs while being more frequent in male authors.

To interpret these instances of thematic "gender-crossing" would require a more thorough analysis than is possible here. And we should be cautious about what we read into these results, as our corpora are small and do not represent balanced selections of text. For instance, some patterns might be due to individual authors or works. However, these observations support our conclusion that the present TM approach leads to sound results: the gender biases relating to characters and authors correspond to each other, but they do not align entirely. Investigations into such discrepancies can provide insight into how literary themes connect to gender in different ways.

Another thing we can see is that gender bias is slightly more pronounced in the Classics corpus, both as regards characters and authors. Such comparisons can also be made for specific themes. For instance, "Reading and writing" has moved from a male-dominated arena in the Classics corpus to a gender-neutral

Themes in the Bestsellers corpus	Topic	Themes in the Classics corpus
Existential reflection, mental states	**1**	***Faces, appearance***
Body parts (***faces***), intimacy, violence	**2**	***Mental states, existential reflection***
Young jargon, ***social interaction*** (vague)	**3**	***Social interaction*** (vague)
Watching, ***dancing, music***	**4**	Encounters in urban milieus
War, nations, global politics	**5**	***Disease, health care.*** Accidents
Outdoors (nature, water, weather)	**6**	Character traits
Fighting, violence. ***Animals***	**7**	***Mental states*** – strong emotions
School, childhood, adolescence	**8**	***Money, trade***, work
Sleep, going to bed, waking up	**9**	***Outdoors (landscape, weather)***
Telephone calls	**10**	***Family*** relations
Interiors	**11**	God, religion, faith
Family (ceremonies)	**12**	***Interiors***
Moving in buildings	**13**	***Reading and writing***
Eating and drinking	**14**	Politics (class and gender) – Strindberg
Police work, investigations, law enforcement	**15**	Titles, rank, formal register
Driving, transportation	**16**	Agriculture, ***animals***
Appearance (primarily women)	**17**	Authority – church, nation
Reading and writing	**18**	***Dance, music***, entertainment
Family circle. ***Disease, health care***	**19**	***Eating and drinking***
Money and trade	**20**	Maritime activities and settings (***water***)

Table 4. Comparison of themes retrieved by TM over the two corpora ($k = 20$, $c = 24$). Themes corresponding to each other in the two corpora in bold italics.

one among the contemporary bestsellers. "Eating and drinking" has in a similar fashion moved from a male-dominated arena in the Classics corpus to a female-dominated one in the Bestsellers corpus (see Figure 4). These results are easy to relate to well-known societal changes that have taken place in the time span between the two corpora.

Although the gendered patterns discussed above are indeed evident, we should not overestimate their magnitude. In the most clearly biased themes (found when $k = 100$), the difference is a matter of a factor of 2, both as regards characters and authors. But most topics do not exhibit a pronounced gender bias. The gendered patterns are overall less distinct than we expected, which is in itself an interesting result and something that should be addressed more carefully in the future. We also expected that an increase in k would generate more gender-biased topics (since larger k-values generate more specific themes in the TM), but such a pattern could not be observed.

4.2 Thematic Comparison between the Two Corpora

If we look at the topics from the two corpora when the TM has been asked to deliver broad thematic categories ($k = 20$, $c = 24$), we find a striking number of thematically related topics. We think that a qualitative analysis which shows that most topics are thematically related to a topic in the other corpus (see Table 4) is justified. So, for instance, eating and drinking, reading and writing, money and trade, and existential reflection and mental states, are four themes which

Classics #8 **pengNN**, **köpaVB**, **betalaVB**, **säljaVB**, **arbeteNN**, **skaffaVB**, **kronaNN**, arbetaVB, **arbetareNN**, **kostaVB**, **prisNN**, **lönNN**. [*money, buy, pay, sell, work, get, crown, work, worker, price, salary.*]

Bestsellers #20 **pengNN**, **betalaVB**, **säljaVB**, **köpaVB**, **kronaNN**, **kundNN**, **företagNN**, **kostaVB**, **tjänaVB**, **prisNN**, **bankNN**, **miljonNN**. [*money, pay, sell, buy, crown, customer, cost, earn, price, bank, million.*]

Classics #19 **drickaVB**, **kaffeNN**, **matNN**, ätaVB, **vinNN**, **middagNN**, **glasNN**, **brännvinNN**, **ölNN**, **brödNN**, bordNN, **dryckNN**. [*drink, coffee, food, eat, wine, dinner, glass, liquor, beer, bread, table, beverage.*]

Bestsellers #14 **kaffeNN**, **drickaVB**, ätaVB, **matNN**, **vinNN**, **koppNN**, **glasNN**, **brödNN**, bordNN, **teNN**, **kokaVB**, **flaskaNN**. [*coffee, drink, eat, food, wine, cup, glass, bread, table, tea, boil, bottle.*]

Table 5. Two topics – "Money and trade" and "Eating and drinking" – of obvious cross-corpus equivalence ($k = 20$ and $c = 24$).

in a clear way surface as topics for both corpora. Table 5 shows the remarkable overlap among the top terms for the topics "Money and trade" and "Eating and drinking". So, themes like these seem to be persistent in Swedish literature, even if the corresponding topics also reflect the historical context.

The topics which seem to capture themes which only appear for one corpus do so in a conceptually straightforward way. Themes like telephone calls, driving, and police investigations are retrieved only for the contemporary Bestsellers. For the Classics corpus, by contrast, some topics capture themes which were more important for the economy and society of the time period it represents, such as agriculture, maritime activity, titles, and religion.

Comparisons of this kind lead to interesting observations, which would be possible to follow up in a deeper qualitative analysis. In this way, the use of TM for the exploration of literary corpora suggests new ways for the literary scholar to track themes, changes in themes, and the decline of themes over extended periods, throughout literary history.

5 Conclusions

In this paper, we have proposed an experimental procedure whose central component is TM to extract themes from Swedish corpora of prose fiction. The topics and the term lists derived from them are the products of algorithms whose output is based on statistical processing of text segments in parallel. Which topics are established depends on patterns of co-occurrence among words. These can be due to both conceptual and factual associations of a more general nature, as well as to more idiosyncratic circumstances, e.g. patterns deriving from specific authors.

Our analysis of the results is tentative and sketchy because of the size of the material and the limited scope of the study. We still wish to conclude that

the experiments are successful in consistently producing easy-to-interpret and conceptually plausible topics. We think that our use of χ^2 (chi square) to rank the terms in the topic word lists facilitate the qualitative interpretation, as it lifts the "core" terms, and – so to speak – highlights the central thematic content.

The method works with both small k values, for the retrieval of broad thematic categories, and with larger k values, in which case we get more specific themes. This makes the method flexible and possible to use for different research questions. We have not systematically investigated different parameter settings for the various modules (other than compared choices for $k \in \{20, 40, 100\}$). This would be a challenging task, because of the number of parameters involved, and the lack of precise criteria for assessing the output. The chunking principles are also in need of further evaluation: For instance, there is a need to compare the consequences of using chunks captured by a window-based procedure with a design based on paragraphs or larger units as chunks.

We would like to highlight three areas in which "distant reading" based on TM can be of value to scholars of literature:

First, the use of TM allows us to investigate themes in a bottom-up fashion, where the computer-generated output is the point of departure for a qualitative analysis. This approach stands in contrast to methods driven by the researcher's preexisting ideas about the nature of given themes. In this way, it prompts us to read familiar literary material in new ways, by giving us condensed pictures of the topics/themes which are quantitatively justified by the textual surface facts. In particular, we hope that we have shown that the use of TM allows us to make plenty of interesting observations about Swedish literature.

A second advantage is that TM with labeled chunks supports analysing topics and their dependence on other properties. In the present study, we performed a gendered thematic analysis which connects thematic content to biases concerning the gender of characters and authors. Our use of pronoun evidence to connect paragraphs to character gender is admittedly a far from perfect operationalization. It could be improved by the use syntactic parsing and retrieval of more fine-grained labeling of literary characters. It is easy to imagine studies in the same vein where the focus would lie on other categories, such as race/ethnicity, class, age, etc.

A third interesting kind of analysis supported by TM is the comparison of different corpora. The collections which we have discussed here, classics from the decades around the turn 1900 and recent titles from the bestseller charts, would hardly invite thematic comparison in traditional literary studies. However, we were able to align the topics derived from the two collections and find remarkable similarities, as well as interesting differences. A related application would be to track how themes develop and recede through time. In this article, this comparison of the topic sets has been a matter of qualitative analysis. A prospect for future work is to develop computational mechanism for aligning topics.

There are also other ways in which TM can serve literary research, which are beyond the scope of this paper. One is to use TM as a component in a search tool,

which assists the scholar in finding occurrences of a given theme in a corpus (see e.g. [32, 17]). Another application to use TM for basic narratological analysis as regards the linear distribution of thematic content in the course of novels, as has been done by Jockers [10]. Performed on a larger scale and combined with broad thematic categories, such analyses can help us find recurring plot structures and allow us to see how narratives work. This is yet another promising idea for future studies.

It is important to remember that the TM output always has to be interpreted qualitatively by the literary scholar in the kind of research we propose. This interpretation is far from a trivial task and requires an understanding of the corpus. As we see it, TM does not provide a method for delivering definite reports about the set of themes inherent in a literary corpus. Rather, the TM output is in need of literary interpretation of a traditional kind to yield new insights and knowledge.

In her book *Reductive Reading* (2018), Sarah Allison develops a closely related point. She argues that one of most important lessons literary historians can learn from computational criticism is that it forces the researcher to lay bare what is taken into account in the analysis and what is not. Such methods are in this way more solid and more transparent than traditional methods for literary research. Allison writes:

> Reductive reading, by contrast, clears space for reading that is *not* reductive. [...] My argument for reductive reading is also a defense of descriptive – call them "weak" – findings: a very strong opening claim can shelter more nuanced claims than a project that open with a more textured or qualified polemic. ([2]: 2–3)

This is similar to how we understand the possibilities of TM for scholars of literature: If we take the caveats of the method seriously and discuss them transparently, TM output can provide a solid quantitative backdrop to further qualitative literary analyses. We consequently reject the idea – that sometimes is loudly expressed in debates – that quantitative and qualitative methods stand in opposition to each other.

Acknowledgements

We would like to thank Litteraturbanken, and its Head, Professor Mats Malm, for making data available to us. We are also grateful to the Disciplinary Domain of Humanities and Social Sciences at Uppsala University for funding.

References

1. Allington, D, Brouillette, S, Golumbia, D (2016), "Neoliberal Tools (and Archives): A Political History of Digital Humanities," *Los Angeles Review of Books*, May 2016.
2. Allison, S (2018), *Reductive Reading: A Syntax of Victorian Moralizing*, Baltimore: Johns Hopkins University Press.

3. Archer, J, Jockers, M (2016), *The Bestseller Code: Anatomy of the Blockbuster Novel*, New York: St. Martin's Press.
4. Barakat, A (2018), "What Makes an (Audio)Book Popular?," master thesis in statistics and machine learning, Department of Computer and Information Science, Linköping University.
5. Bergenmar, J (2017), "Har feminism och vithetskritik en plats inom digital humaniora?," in Erixon, P-O, Pennlert, J (ed.) *Digital humaniora – humaniora i en digital tid*, Göteborg: Daidalos, pp. 77–97.
6. Berglund, K (2017), "Killer Plotting. Typologisk intriganalys utifrån fjärrläsningar av 113 samtida svenska kriminalromaner," *Tidskrift för litteraturvetenskap*, (3–4), pp. 41–68.
7. Blatt, B (2017), *Nabokov's Favourite Word is Mauve: The Literary Quirks and Oddities of Our Most-Loved Authors*, London & New York: Simon & Schuster.
8. Blei, DM (2012) "Topic Modeling and Digital Humanities," *Journal of Digital Humanities*, Vol. 3, No. 2.
9. Goldstone, A, Underwood, T (2014), "The Quiet Transformation of Literary Studies: What Thirteen Thousand Scholars Could Tell Us," *New Literary History*, vol. 45, pp. 359–384.
10. Jockers, Matthew (2013) *Macroanalysis: Digital Methods and Literary History*, Urbana: University of Illinois Press.
11. Jockers, M, Mimno, D (2013), "Significant Themes in 19th-Century Literature," *Poetics*, vol. 41(6), pp. 750–769.
12. Kokkinakis, D, Malm, M (2015), "Detecting Reuse of Biblical Quotes in Swedish 19th Century Fiction using Sequence Alignment," *Corpus-based Research in the Humanities workshop (CRH)*, pp. 79–86.
13. Limin Yao, L, Mimno, D, McCallum, A (2009) "Efficient methods for topic model inference on streaming document collections." In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '09)*. ACM, New York, NY, USA, 937–946. DOI: https://doi.org/10.1145/1557019.1557121.
14. *Litteraturbanken*, "The Swedish Literature Bank," URL: https://litteraturbanken.se/om/english.html.
15. Liu, A (2012), "Where is Cultural Criticism in the Digital Humanities?," in *Debates in the Digital Humanities*, Matthew K. Gold (ed.), Minneapolis: University of Minnesota Press, pp. 490–509.
16. Lundblad, K (2017), "Digital humaniora – en pleonasm i den digitala kulturen," in Erixon, P-O, Pennlert, J (eds.) *Digital humaniora – humaniora i en digital tid*, Göteborg: Daidalos, pp. 27–51.
17. Magnusson, M, Öhrvall, R, Barrling, K, Mimno D (2018), "Voices From the Far Right: A Text Analysis of Swedish Parliamentary Debates," working paper, SocArXiv Papers, April 2018, URL: https://osf.io/preprints/socarxiv/jdsqc/.
18. Mandell, L (2015), "Gendering Digital Literary History," in Schreibman, S, Siemens, R, Unsworth, J (ed.) *A New Companion to the Digital Humanities*, Chichester: Wiley, pp. 511–523.
19. Manning, CD, Raghavan, P, Schütze, H (2008) *Introduction to Information Retrieval*, Cambridge University Press.
20. McPherson, T (2012), "Why Are the Digital Humanities So White? or Thinking the Histories of Race and Computation," in *Debates in the Digital Humanities*, Gold, MK (ed.), Minneapolis: University of Minnesota Press, pp. 139–160.
21. McCallum, AK (2002) MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

22. Moretti, F (2000), "Conjectures on World Literature," *New Left Review*, vol. 40(1), pp. 54–68.
23. Navarro-Colorado, B (2018), "On Poetic Topic Modeling: Extracting Themes and Motifs From a Corpus of Spanish Poetry," *Frontiers in Digital Humanities*, vol. 5(June 2018), URL: https://doi.org/10.3389/fdigh.2018.00015.
24. Newman D, Asuncion, A, Smyth, P, Welling, M (2009) "Distributed Algorithms for Topic Models." *Journal of Machine Learning Research* 10 (December 2009), pp. 1801–1828.
25. Norén, F (2016), "Information som lösning, information som problem: En digital läsning av tusentals statliga utredningar," *Nordicom Information*, vol. 38(3), pp. 9–26.
26. Norén, F, Snickars, P (2017), "Distant reading the history of Swedish film politics in 4500 governmental SOU reports," *Journal of Scandinavian Cinema*, vol. 7(2), pp. 155–175.
27. Rhody, LM (2012), "Topic Modeling and Figurative Language", *Journal of Digital Humanities*, vol. 2(1), pp. 19–35.
28. Roe, G, Gladstone, C, Morrissey, R (2016), "Discourses and Disciplines in the Enlightenment: Topic Modeling the French Encyclopédie," *Frontiers in Digital Humanities*, vol. 3(January 2016), article 8, URL: https://doi.org/10.3389/fdigh.2015.00008.
29. Schöch, C (2017), "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama," *Digital Humanities Quarterly*, vol. 11(2), URL: http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html.
30. Steyvers, M, Griffiths, T (2007) "Probabilistic topic models." In Landauer, TK, McNamara, DS, Dennis, S, Kintsch, W (Eds.), *Handbook of latent semantic analysis*, Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers, pp. 427-448.
31. *Svensk Bokhandel*, "Top lists", URL: http://www.svb.se/toplists.
32. Tangherlini, T, Leonard, P (2013), "Trawling in the Sea of the Great Unread: Sub-corpus Topic Modeling and Humanities Research," *Poetics*, vol. 41(6), pp. 725–749.
33. Underwood, T (2014), "Theorizing Research Practices We Forgot to Theorize Twenty Years Ago," *Representations*, vol. 127(1), pp. 64–72.
34. Östling, R (2013) "Stagger: an Open-Source Part of Speech Tagger for Swedish." *Northern European Journal of Language Technology*, Vol. 3, pp. 1–18 (2013).