TECHNISCHE UNIVERSITÄT DARMSTADT
DIGITAL HUMANITIES COOPERATION
UNIVERSITÄT KONSTANZ

Thomas Weitin, Katharina Herget

Falcon Topics

Pamphlet #4
November 2016

Ed. Thomas Weitin

TECHNISCHE UNIVERSITÄT DARMSTADT
DIGITAL HUMANITIES COOPERATION
UNIVERSITÄT KONSTANZ

Thomas Weitin, Katharina Herget

Falcon Topics

Abstract

Topic modeling is one of the more promising quantitative procedures for exploring semantic structures. The creators of the corresponding algorithms use large text sets to investigate hidden thematic connections which cannot be perceived by the eye alone.

We have, by contrast, tested a medium-sized corpus of novellas that can be explored using both individual readings and statistical procedures. We were motivated by a previously little considered observation: The scholars who have been able to implement statistical procedures for literary corpus analysis in a pertinent way were all extraordinarily familiar with their respective corpora.

Because we were also very familiar with our set, it was possible for us to order the topics being studied according to text-relevant themes. Using the keyword "falcon topics," we describe another type of Topics, ones which seem to reflect the special character of individual texts.

ISSN: 2629-7027

Ed. by Thomas Weitin

Katharina Herget, Thomas Weitin

Falcon Topics

Semantics remain one of the greatest challenges for digital literary analysis. Whenever a person reads, he or she proceeds in most cases with an intuitive orientation toward meaning. He or she expects meaning and attempts to understand the significance of what has been read. Professional literary instruction in schools or universities often has to first thwart this hermeneutic intuition, in order to be able to let an analysis take place prior to interpretation. At the same time, hasty types of readings can provide a clue as to how well our brain feels with meaning. For the computer, in contrast, the handling of meaning is a big problem, because it regularly proves too complex to be operationalized and calculated. Topic Modeling is among the procedures of distant reading that aim to calculate meaning.It is concerned with statistical models whose algorithms render latent semantic structures visible in large text corpuses. It investigates groups of words which are likely to appear together. From these "Topics" we learn about the themes of the texts being analyzed, about the figures which appear in them, as well as about plots and their settings. Without ourselves actually reading, we gain insight into connections within the content. David Blei, whose writings have been important for popularizing Topic Modeling in the Digital Humanities, makes the classical too-big-to-read argument in order to mark the procedure's range of application: this type of approach studies large text quantities that surpass human reading abilities with a view toward hidden thematic information that otherwise remains invisible to the eye of an individual reader.[1] Only with the appropriate distance, with the statistical analysis of very many texts, do these "hidden structures"[2] emerge.

Topic Modeling has most widely been practiced in the quantitative social sciences, primarily in political science. Here we find types of texts and corpora, for instance, parliamentary speeches or the protocols of executive or legislative proceedings, that lend themselves well to this type of analysis.[3] A

[1] „While more and more texts are available online, we simply do not have the human power to read and study them [...]. To this end, machine learning researchers have developed probabilistic topic modeling, a suite of algorithms that aim to discover and annotate large archives of documents with thematic information." (David Blei: Probabilistic Topic Modeling. In: Communications of the AMC 55.4, p. 77)

[2] Ibid., p. 79.

[3] In a large scale project, the Swiss Federal Archive in Bern, for instance, is currently assessing the Topics of meetings of the Federal Council, of the Swiss government. And even the political agendas of the European Council are being studied using Topic Modeling. (see Derek Greene, James P. Cross: Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach).

second emphasis is the analysis of digital communication, for instance, emails or social media such as Facebook and Twitter. Here, too, the text corpora offer good fodder for Topic Modeling algorithms. Two of the three most popular Topic Modeling packages for statistical software R come from researchers who work in the context of Facebook.[4] For digital literary studies, Topic Modeling holds great promise. At the beginning of the renaissance of quantitative procedures in philology, it was stylometric analyses in particular that were able to catch the attention even of colleagues who work with more traditional methods.Because these analyses are mostly based on the frequency of words, critics often had the impression that digital literary studies is merely a matter of counting words and that it is incapable of getting to that which mostly interests readers, namely, to the content and meaning of texts. Hence, in his book Macroanalysis, Matthew Jockers begins the chapter on Topic Modeling with the suggestion that the semantic-oriented method could ultimately grasp "plot, character, and theme."[5]. Nevertheless, that chapter is exclusively focused on "themes" classically studied by Topic Modeling. He demonstrates that the procedure is appropriate for the analysis of both the individual text as well as the corpus. By undertaking comparative corpus analysis, Jockers is able to use Topic Models to classify texts according to the variables "gender" and "nationality."The results show what one perhaps had feared from "macro trends," namely, that clichés come to light: Men most like to write about "pistols and other guns," women about "female fashion," Scots are preoccupied with their "dialect," Americans with "us dollars and us cities," and the Brits naturally with "hounds and shooting sport."[6] Jockers points to the significance of individual outliers (no less than 33 percent for nationality and 14 percent for gender), yet he does not consider them in the context of the applied method. In more recent studies, he concentrates his thematic analyses very strongly on the field of emotion, although he replaces the Topic Modeling[7] used in *Macroanalysis* with a dictionary-based Sentiment Analysis.[8]

What exactly can and do we want to discover using Topic Models? On the level of the individual text the answer seems to be as clear as with large corpora. In the individual text we can, for example, visualize which themes are dominant in which chapters of a novel. And in large corpora we can make hidden thematic structures visible. In the former field of application, the method accomplishes something that we could also achieve by reading; in the latter, the method takes effect where the

[4]Namely, Johnathan Chang and David Blei himself (see Matthew L. Jockers: Text Analysis with R for Students of Literature. Cham u.a. 2014, p. 136).

[5]Matthew L. Jockers: Macroanalysis. Digital Methods & Literary History. Urbana et al. 2013, p. 118.

[6]Ibid., p. 149-153.

[7]Cf. Ibid., p. 136.

[8]With his „syuzhet" package for R, Jockers initiated an intense Internet discussion. (see: Annie Swafford: Problems with the Syuzhet Package, Blogpost from 2 March 2015 https://annieswafford.wordpress.com/2015/03/02/syuzhet/

individual reading is insufficient, and replaces it. It is of course possible to evaluate a distant-reading result from Topic Modeling through close reading, in order, for example, to identify that a particular male author from Jockers' corpus tells penny dreadfuls in accordance with the macro trend, whereas a female author, deviating from the trend, makes use of other themes than "female fashion." With the use of statistical methods in literary studies, two gestures of thought frequently appear in tandem: One is happy to show that a method functions, which usually works by confirming an already existing finding or by arriving at a result that is predictable given the text corpus – Americans really do have different topics than the English! And at the same time, the research explains that it is precisely the deviations from the macro trend that are interesting.

Statistically, the outliers belong so reliably to the macro trend as the standard deviation does to the normal distribution of the data set. In terms of literary studies, the pair of thought gestures is reminiscent of the old hermeneutic wish for a reciprocal understanding of part and whole. Yet it is not enough to find the outliers interesting. And if while reading an individual text that deviates statistically I claim to discover reasons for this deviation, the question can always be asked as to whether the individual hermeneutic finding and the result of statistical corpus analysis have anything to do with one another in the first place. Let us imagine that I take from Jockers' corpus the text of a woman author which very starkly contradicts the measured Topic trends for female writers. And I find out that this author was a physicist, who incidentally had nothing to do with "female fashion." In this case, then, the interpretative connection between individual historical understanding of the text and Topic Modeling of the corpus would be correct. There is nothing wrong with corroborating this sort of a thesis related to an individual text using quantitative evidence, especially since the handling of all quantitative procedures requires much practice, which is most strongly fostered by concrete questions. Whoever is well practiced will seek to deploy Topic Modeling in ways that are no longer merely descriptive, but also explorative. We do not only want to take a look at Topic-participation and Topic-distribution; rather, we want to thereby discover something about connections and relations between texts and authors with their respective manifold characteristics. We are concerned with the connections within the modeled corpus and beyond.

While investigating such connections, it helps tremendously to know something about the texts and authors, even if we can not read each one of the texts. To this extent, the initially so convincing too-big-to-read argument is somewhat misleading. From the perspective of Facebook analysts, the emergence of hidden structures out of data sets that can only be assessed statistically might provide the correct epistemological role model. Literary scholars, on the other hand, have a different knowledge

praxis. In most cases, we study corpora about which we know a great deal, whether through the knowledge of certain texts, through knowledge of historical context, or through systematic knowledge about genres, forms, intertextuality, and so on. Those scholars who have succeeded in deploying statistical procedures for literary corpus analysis in a pertinent way were all extraordinarily familiar with their respective corpora. This is the case, for example, for John Burrows, who tested the Delta procedure, named after him, for determining the stylometric distance between texts using a corpus with 25 English authors of the seventeenth century; or for Matthew Jockers, who, thanks to his outstanding knowledge of this particular literary history, was able to make his corpus of Irish-American literature into an ideal object of study in *Macroanalysis*.

In the LitLab in Darmstadt, we have explored Topic Modeling using the corpus of the *Treasure Trove of German Novellas (Deutscher Novellenschatz)*, a collection of 86 novellas edited by Paul Heyse and Hermann Kurz. The 24 volumes of this collection appeared between 1871 and 1876. We digitized them, processed them as a corpus in TXT and TEI-XML-Standard (corrected OCR), and enhanced them with metadata for the individual texts and their authors. In the process, we learned a great deal about our corpus. We have a genre anthology on our hands that defines itself as a "sample compilation" (Mustersammlung)[9] with a clear consciousness of the epoch of realism. 49 of the 86 novellas come from the period after 1848, the two oldest from 1811, and the youngest from 1875. Almost strictly individual texts have been collected; only four authors can lay claim to two texts in the collection. 12 female authors are juxtaposed to 70 male authors.

One advantage that we have come to cherish while working on this corpus is its mid-size. It is still small enough for our individual reading competency, but also large enough to warrant statistical analysis. We continually gather new data and metadata in a large Excel spreadsheet, although the impulses for this frequently come from our readings of individual texts. Alongside the novellas, relevant texts of literary history from the context of the *Treasure Trove of Novellas* play an important role when it is a matter of understanding the function of the collection as an instrument of literary historiography. For example, we consult the introduction to the first volume of the *Treasure Trove of German Novellas* written by Paul Heyse. There, Heyse sketches a realist poetics of the novella genre, which seeks to identify the relationship between theme and form as the genre's distinguishing feature. With a view to the common demarcation from the novel, Heyse dismisses "length" (*Längenmaß*) as an insufficient criterion; rather, something has to reside "in the theme" that "by necessity pushes to the one form or

[9]Paul Heyse, Hermann Kurz: „Einleitung". In: Deutscher Novellenschatz. München 1871, p. 24.

the other."[10] Heyse believes that "such a simple form" as the novella is "not [appropriate] for every theme of our friable modern cultural life."[11] However, instead of isolating the jurisdiction of the genre's content, he distinguishes the novella from the novel by identifying special heuristic procedures for the reduction of complexity that are unique to the novella.[12] While the novel, as a genre of reflection, illuminates its topics "in an exhaustive manner from all sides," the novella "condenses" everything "in one point."[13] The novella's thematic treatment is based on the "isolation of the experiment," which means in poetological terms to concentrate on a fundamental motif and to shape "a stark silhouette." We can best test whether that works, according to Heyse, in an "experiment," namely by probing whether it is possible "to summarize the content in only a few lines."[14]

This empirically-oriented poetics of genre has gone down in literary history as the "falcon theory," because Heyse explicates his test using the novella by Boccaccio in which a falcon sacrificed for love contains the condensing quality of a fundamental motif according to which the novella's plot can in fact be summarized in short sequences. The "falcon text" is given as a central criterion of selection after the texts were chosen for the German Treasure Trove of Novellas. Writers wishing to produce novellas were expressly instructed to submit their material to this test first and to ask "where is 'the falcon'"?[15]

For the Topic Modeling of the *Novella Treasure Trove* we used the R package MALLET within R-Studio.[16] Because with Topic Modeling we systematically search for connections among the content, the corpus first has to be purified of function words and such words bearing information whose frequency would dominate the results (for example: "have"). We adjusted an existing stop word list for German to our purpose, accounting above all for a large number of the names that had been very strongly present in the Topics during the first analysis tests with the corpus. Topic Modeling is based on two premises:[17] There is a certain number of commonly used words whose common appearance repeats itself regularly in texts like a pattern. Those are the Topics. (1) Every single text in a corpus can be described in terms of how strongly every one of these Topics is present in it and which words belong to it. (2) In accordance with this presupposition, the algorithm works with a bag of words containing

[10] Ibid., p. 17.
[11] Ibid., p. 20.
[12] Cf. Thomas Weitin: Heuristik der Novelle. In: Albrecht Koschorke et al. (Eds.): Komplexität und Einfachheit. Stuttgart 2017 (in preparation).
[13] Paul Heyse, Hermann Kurz: „Einleitung", p. 18 (first quotation), p. 17.
[14] Ibid., p. 18 (first quotation), p. 19.
[15] Ibid., p. 20.
[16] The following analyses refer back to the Master's thesis by Katharina Herget: Die Literaturgeschichtsschreibung des Deutschen Novellenschatzes Paul Heyses: Qualitative und quantitative Perspektiven. Typescript. Konstanz 2015.
[17] Cf. Blei, p. 77f.

all of the remaining words in the corpus and distributes them randomly along the previously defined number of topics until a common appearance has stabilized. The observed patterns are then collated with the actual emergence of the Topics in the texts. In other words, what gets measured are the Topics, the Topic-parts in the text, and whether words in a text belong to the Topics.

The corpus premises, which have to be defined using the MALLET package and whose definition represents the respective model, are shaped in correspondence with those with which the experiment is working. If the stop word list has been identified, it is a matter of optimizing three parameters: the number of Topics, the number of words per Topic, and the number of iterations until the hoped for stabilization of results. A model is stable when the co-occurring word groups, which it has produced as Topics, can be repeated. At issue is the reliability of the model. In our experiments with the 86 texts of the novella treasure trove corpus, a model with 100 Topics, 10 words per Topic, and 10000 iterations produced relatively reliable results. 57% of the Topics proved to be reproducible – though not immediately. In order to attain this, we had to make an essential alteration in the experiment. Despite great effort with many different settings of the three parameters (number of Topics, words per Topic, and iteration), we were not able to reach a reproducible result on the basis of the 86 individual texts of our collection. To do so, we first had to prepare the corpus so that the algorithm no longer treated the individual text as a fundamental unit, but instead processed automatically generated chunks of 300 words each, totaling 5,380 individual documents. After such pre-processing, our model was then relatively stable.

It is well known that the novella's basic formula is 2 + 1: Two are made for each other, a third interrupts, a conflict erupts with a good or bad ending. Hannelore Schlaffer, drawing on Gottfried Keller, has called this "novella mathematics" (*Novellenmathematik*).[18] We knew from individual readings that the novella treasure trove contains an overwhelming majority of such marriage novellas. If we include marriage novellas which tell not about the path to marriage but about married life afterwards, the proportion is more than 90

Heirat

12	müllerin heirathen jurancon verwandtschaft base müller sich's stiefvater verlobung schwager
13	onkel großmutter julie oberförster brink trauung liebes berger brigitten vollends
17	wirthin baron gegend thurn freifrau braut lich fräulein lothar schlosse
22	herzog reise advocat durchlaucht baron studenten polen kurland bursch hochzeit
40	mädchen signora urballa burschen signor heirathen villa padrona signoria baldo

[18] Cf. Hannelore Schlaffer: Poetik der Novelle, Stuttgart 1993.

Based on our readings, however, we had expected that this absolutely dominant theme would emerge even more prominently.[19] Nevertheless, marriage is in most novellas a major theme. We see of course that our Topics have blemishes: "sich's" should have been included in the stop words, "lich" is obviously a mistake, and a few names have also remained. A comparison shows that, semantically and in terms of the gathered types of words, the Topics are pertinent in different ways for the theme of marriage. Topic 12 is certainly relevant with "engagement" (verlobung), "kin" (Verwandschaft), "brother-in-law" (schwager) and "marry" (heirathen). As is Topic 40 with "girls" (mädchen), "boys" (burschen), and "marry" (heirathen). The other three Topics each contain only one relevant substantive (13: "marriage" [trauung], 17: "bride" [braut], 22: "wedding" [hochzeit]), 17 and 22 are purely substantive Topics.

If we decide to treat these five Topics as marriage Topics, we can compare them with other thematic clusters that can be grouped into 100 Topics from our model. The thematic areas of military, religion, law, arts, wealth and farmers can be classified into at least three Topics.

Militär

20	offizier graf nacht pferde offiziere pferd husar junge husarenoffizier hauptquartier
66	chef lieutenant freund ehre freunde gefährten ring flagge offiziere freundes
76	major haide pferde ritten herum uwar maroshely gitter ritt lande
78	teufel general soldaten könig hauptmann commandant offizier dienst oberst wache
98	franzosen krieg frankreich könig männer vaterland volk kaiser feind französischen

Religion

15	gott ehre grab alte schwert herzog gnade begraben schönen letzten
60	heiligen nonne bild kloster antlitz kunst kirche malen bildes entgegnete
76	major haide pferde ritten herum uwar maroshely gitter ritt lande
91	vermittelung gemeinde klöster gesunden gottlosen zustand dünnen gebilde gesunde eingebüßt

Justiz

10	silberburg rothenburg stadt scharfrichter wolf heyliger thal römerhöhe garten rath
21	müller mühle schmied justizrath actuar gerichtsdiener knabe meister schmiede müllers
70	frei schuldig staatsanwalt rothmann geschworenen verhandlung schwiegersohn angeklagten männer vorsitzende
92	richter stadtschreiber galgen gericht urtheil herren rath strafe freiheit zuletzt

[19] The absolute frequencies of relevant marriage concepts in the corpus increased the expectation created by the reading impression. Of the approximately 1.6 million tokens of our entire corpus, at least 1,309 come to a very narrowly defined marriage thesaurus: hochzeit [wedding]* (260), braut [bride]* (333), heirat [marriage]* (515), trauung [wedding]* (37), bräutigam [groom]* (125), vermählt [wedded]* (39).

Künste

43	brief las schrieb lesen briefe schreiben geschrieben buch papier gelesen
45	novelle geschichte namen form roman dichter stoff meister erzählungen krieges
47	alte bild maler kunst künstler freund bilder gemälde fremde junger
54	kennen natur lernen nennen bildung kunst geist meisten höheren erscheint
83	lied musik singen spielen sang talent spielte geige klavier sängerin
100	tanz tanzen tanze vicomte tanzte jungen musik mädchen musikanten altenkreuz

Vermögen

50	schatten gold herrn grauen sonne tasche goldes seckel pracht vieles
61	geld gulden thaler handel summe wasser schulden littauer wirth verkauft
67	ring finger musik eltern palast schöne niemals kästchen stehen gewaltigen
86	fräulein sprach schmuck meister könig arbeit that kästchen geheimniß paris
96	kästchen gesellschaft wagen zimmer schöne beutel geld wodurch leben vergnügen

Bauern

8	bauer gott großmutter alte magd kind heiligen knabe worte seele
68	jäger bauern mühle förster bauer wald thal hütte jüngling felsen
74	acker kinder bauern puppe nase steine jahren wilde aecker loch

This overview shows that the marriage Topics are about as pronounced as the other thematic areas are. Our impression from reading, namely, that the marriage theme is absolutely dominant, has hence not been corroborated. And yet the overall picture of the Topics does indeed correspond to what we know about our corpus and its individual texts. Marriage and starting a family are mostly connected with questions of wealth in realist novellas. It is the life of simple men and women (farmers) that is depicted in the corpus of well-represented village stories. Religion plays a large role; there is a strong tension between superstition and secularization. Many conflicts are also handled by the court of law. At the same time, the stories are about the decline of the nobility and the rise of the bourgeoisie. Military careers can be found on all social levels, as they are also a matter of social mobility. There are individual artist novellas, but the strong presence of this Topic is rather surprising for readers.

This overview of Topic Modeling, reading, and knowledge of literary history could lead us to conclude that our model has functioned. The statistical Topics produced by the algorithm are relatively consistent and we can classify them into relevant themes. Our historically-informed interpretation of content and the statistical corpus analysis seem to correspond to one another.

But is this in fact the case? Or, to pose the question even better: How appropriate is it to classify and interpret the results of a statistical Topic Model in this hermeneutic way? And where does that lead us? Let us consider an example. We ascribed the following Topics to the thematic area of wealth:

67 ring finger musik eltern palast schöne niemals kästchen stehen gewaltigen
86 fräulein sprach schmuck meister könig arbeit that kästchen geheimniß paris

Although "ring" (ring) and "palast" (palace) in Topic 67 and "schmuck" (jewelry) with "könig" (king) in Topic 86 could each be understood in combination with "kästchen" (casket) as indicators of wealth, our classification decision might have been influenced by a different Topic. In Topic 96, "kästchen" (casket) appeared together with the relevant noun "geld" (money):

96 kästchen gesellschaft wagen zimmer schöne beutel geld wodurch leben vergnügen

As we already determined with the marriage Topics, we could say that individual Topics are relevant for a theme in different ways. How we undertake such classifications, is ultimately a question of interpretation.

As we discussed the order of our Topics, we had, along with the literary history of realism, two things in the back of our minds. We thought about what the Topic Modeling algorithm calculated: Topics in the corpus, Topic proportions in the text, and the belonging of words in a text to the Topics. And we thought about the editors of the novella treasure trove who only wanted to include texts in the collection that passed the "falcon test," in other words, only ones with a strong fundamental motif, according to which the content can be readily summarized. Because we were familiar with the texts of our corpus and certain names and designations of place remained, we could immediately think of a matching novella for some of the Topics. In these cases, we began to test whether in fact all words of a respective Topic could be found in the given novella. The result was astonishing: For 32 of the 100 Topics, all 10 Topic words did indeed stem from one and the same novella. Goethe's *New Melusine* (*Neue Melusine*), the novella placed by the editors at the beginning of the *Treasure Trove* because Heyse considered it an indispensible style template, is the only one which even forms two entire Topics.

We were confused. In order to attain reliable results, we dissolved the unity of novellas and instead used 5,380 chunks of 300 words each to determine the Topics. Was it not highly improbable that in almost a third of the cases the words of a Topic would be those of a particular text? Of course, we had right away one name for our observation. In honor of Paul Heyse, we began to refer to Falcon Topics. Did they form a semantic "silhouette" in the way that Heyse had imagined it? Could we discern the characteristic plot of the novella from these Topics? Let us take a look at a couple of examples:

Falcon Topics

Tieck: Des Lebens Überfluss (Life's Abundance)
65 treppe polizei holz ofen stufen lachen emmerich sieh niemals feuer

Goethe: Die neue Melusine (The New Melusine)
67 ring finger musik eltern palast schöne niemals kästchen stehen gewaltigen
96 kästchen gesellschaft wagen zimmer schöne beutel geld wodurch leben vergnügen

E.T.A. Hoffmann: Das Fräulein von Scudery (Mademoiselle de Scudéri)
86 fräulein sprach schmuck meister könig arbeit that kästchen geheimniß paris

Kleist: Die Verlobung in St. Domingo (The Betrothal in St. Domingo)
72 fremde neger fremden alte familie knaben herrn weißen bette babekan

Schmid: Mohrenfranzel (Moor Franzel)
89 königin spielen burschen stück saba bursche schiffer schwarze bühne zunftmeister

Roquette: Die Schlangenkönigin (The Queen of Snakes)
48 kahn wasser bild schlangenkönigin dorfe schlangen ufer entgegen gegend leipe

In Topic 65, we can immediately recognize the "falcon" of Tieck's novella *Life's Abundance*, in which a couple, defying economic necessity, burns the "stairway" to the outside world in order to retreat entirely into romantic togetherness. Topics 67 and 96 both contain the "casket" (*Kästchen*), in which the miniature world of the beautiful dwarf princess in Goethe's *New Melusine* is transported along with the palace. A "casket" is also the thing symbol in Topic 86. For E.T.A. Hoffmann, it transports the jewelry of the goldsmith René Cardillac that is highly calamitous for its noble recipients (Mademoiselle de Scudéri). Kleist's Haiti novella *The Betrothal in St. Domingo* shows its historical-dramatic "silhouette" just as much through the conflict surrounding race and the "stranger" as Hermann Schmid's novella *Moor Franzel* does through a traveling stage operated by mariners in the Bavarian province, whose only black actress plays the Queen of Saba. Finally, the "snakes" in Otto Roquette's Spree Forrest novella *The Queen of Snakes* expresses the misogynous metaphor of sexuality in a nutshell.

Whoever believes that the novellas which form such Falcon Topics really do stand out within the corpus and are distinguishable in a special way is tempted to impose further explanatory burdens on the statistical phenomenon. For instance, in connection with the observation that four of five Topics in the thematic area of wealth constitute Falcon Topics, whereas there are no Falcon Topics at all among the themes of religion and arts. A closer look at the corresponding Topics seems to confirm

the hermeneutic impression. In contrast to the prominent falcons in the area of wealth, the following examples from religion and arts do not invoke any particular text to even the informed reader:

81	kirche pfarrer sonntag heilige messe gemeinde priester geistliche heiligen predigt
45	novelle geschichte namen form roman dichter stoff meister erzählungen krieges

However, we want to curb speculation here and instead attempt to determine whether the presumed finding is rather a systematic mistake or at least the result of the manipulation of the parameters of our model. One thing to consider would be that 100 Topics are simply too many for the size of our corpus, so that the words appearing together are therefore to be found in one and the same text.[20] If we require less Topics from the algorithm, following this line of thought, the likely appearing word patterns will be distributed over more texts and the Falcon Topics will disappear. We reduced the number of Topics step-by-step, but we did not find confirmation of this presumption. To be sure, the Falcon Topics decrease with the sinking number of Topics, but they do not disappear entirely. If we had 32 for 100 Topics, we nevertheless find 2 among 20 Topics. The 20-model yields the following Topics:

1	frau vetter munde conrectorin vetterchen that schäferle stadt general buchenberg
2	mynheer see chef schiff boot lieutenant meister zorgenhof freimeister amsterdam
3	hunde bloß major herren pferde dachte notar junker koppigen wilde
4	könig königin herzog stadt ehre soldaten franzosen offizier königs könige
5	schatten ring kästchen fräulein schmuck arbeit könig meister paris arzt
6	wagen pferde pferd kutscher dame klarinett reiter graf burschen signora
7	pfarrer kirche berg bauer pfarrerin großmutter burg magd schwester knabe
8	sonne wasser erde himmel luft wald straße garten schatten land
9	zeit ließ hause tage jungen tag wußte fand that tochter
10	müller küster mühle schmied stadt justizrath silberburg herrn rothenburg hedeper
11	landrichter jäger lappen förster felsen fjord kaufmann missionär thal hütte
12	herr marquis doctor herrn hahn stadt professor todten gast fräulein
13	rief vater alte herr mann frau fort sagen sehen mädchen
14	hand augen stand schien nacht gesicht hielt trat lag weg
15	dorf arbeit bauern hof juden jude littauer feld acker wasser
16	wein sprach tisch glas trinken gäste trank keller herren rose
17	buch prinzessin geschichte london novelle werth bildung wirklichkeit deutsche gelesen
18	bild kunst maler nonne künstler musik heiligen spielen bilder bilde
19	leben liebe herz welt menschen seele glück herzen fühlte thränen
20	graf frau baron kind vater fräulein gräfin grafen marquis frug

If we compare the 20-model with the 100-model,[21] systematic commonalities stand out. In both, we find Falcon Topics (red) as well as thematic Topics (green). The proportion of Falcon Topics lessens, as we said, by around a third (32/100) to 10 percent (2/20), whereas the proportion of thematic Topics

[20] We thank Sabine Bartsch (Darmstadt) for calling our attention to this idea.
[21] Cf. Appendix, p. 17ff.

lies at around 40 percent in both models. We rediscover in the drastically reduced model not all of our themes, but indeed exclusively those that also appear in the much larger model, for example, Topics pertaining to the themes of art (17 and 18), religion (7) and nobility/military (4, 20) discussed above, as well as to nature (8), eating (16) and love (19). We had seen that the corresponding thematic classifications are a matter of interpretation. Of greater systematic importance is the conclusion that semantic coherency within the Topics did not improve by reducing the number of Topics. It stays just as good, and we see once again that thematic Topics can be relevant for their theme, or, even better, that they express it more or less coherently. The model which is, in terms of the systematicity of its results, very similar yet more compact makes comparisons easy for us. Let us pursue this observation a bit further and compare the thematic Topics of the 20-model to one another. With a view to the Topics 16 and 19, we might be tempted to believe that coherency emerges whenever a Topic contains relevant verbs alongside substantives.

16 wein sprach tisch glas trinken gäste trank keller herren rose
19 leben liebe herz welt menschen seele glück herzen fühlte thränen

Here, semantic coherency seems to be an effect of possible syntactic connections that are triggered by the co-occurrence of different types of words. For both Topics, it is easy to imagine scenes that could be described in sentences that can be built out of the words of the Topics. However, a comparison with Topic 8 teaches us that even purely substantive Topics can invoke relevant scenes. For the concepts of Topic 8, we can easily imagine a walk through nature.

8 sonne wasser erde himmel luft wald straße garten schatten land

Alongside our distinction between Falcon Topics and thematic Topics, we also marked two additional groups of Topics in our 20-model in grey and yellow. Each of these Topics is systematically close to one of the types of Topics, yet does not belong to any of them. The ones marked in yellow gave us reasons to treat them as falcons due to certain words, in particular, names, proper names and designations of location. However, a test showed that they do not only consist of words from one and the same text, even if the proportion of words, for which that was the case, is considerable in each case. The Topics marked in grey appear to be very coherent semantically, without really being thematic. Let us take a closer look at this Topic group:

9 zeit ließ hause tage jungen tag wußte fand that tochter
13 rief vater alte herr mann frau fort sagen sehen mädchen
14 hand augen stand schien nacht gesicht hielt trat lag weg

Without further ado, we could build meaningful sentences out of these Topics consisting of verbs, substantives, and adjectives, and they also evoke certain scenes for us. Assigning them to a theme, however, would in each case require a strong interpretation, which for at least Topics 9 and 13 would not be justified on the basis of the Topic alone. Topic 14, on the other hand, with the nightly scenery it invokes, could potentially be assigned to the thematic Topics. It is a liminal case.

Let us now take a look at our two remaining Falcon Topics. Topic 5 is an old acquaintance. All of its words derive from E.T.A. Hoffmann's *Mademoiselle de Scudéri*, which already stood out as a falcon in the 100-model.

E.T.A. Hoffmann: Das Fräulein von Scudery (Mademoiselle de Scudéri)

Model with 100 Topics
fräulein sprach schmuck meister könig arbeit that kästchen geheimniß paris

Model with 20 Topics
schatten ring kästchen fräulein schmuck arbeit könig meister paris arzt

7 Topic words are stable in the comparison of the two models. And both variants reveal to us the Falcon Topic as one that is typical of its kind: It contains references to figures (René Cardillac is constantly referred to as "master" in the novella, Scudéri as "mademoiselle" (*Fräulein*), and not only in the title), to plot locations ("Paris"), and to objects which are characteristic for the plot ("jewelry" [*Schmuck*], "casket" [*Kästchen*]). Topic 15 is the second falcon of the 20-model. All the words of this Topic come from Ernst Wichert's novella *Ansas und Grita* (*Ansas and Grita*), which in contrast to Hoffmann's novella, did not yet form a Falcon Topic in the 100-model.

Ernst Wichert: Ansas und Grita

Model with 20 Topics
dorf arbeit bauern hof juden jude littauer feld acker wasser

While the geographical designation "Lithuanian" (*Littauer*) is typical for a Falcon Topic and for someone with the corresponding reading knowledge immediately calls to mind the novella in question, the coherent manifestation of the farmer theme is entirely unusual in comparison with the other Falcon Topics. Compared with the Topics classified in the 100-model under the theme "farmers," the Topic of the Wichert novella contains even significantly more relevant content words. This Topic is thus at once a falcon and thematic. To illustrate, here are once again the Falcon Topics from the 100-model:

Falcon Topics (Model with 100 Topics)

Tieck: Des Lebens Überfluss
65 treppe polizei holz ofen stufen lachen emmerich sieh niemals feuer

Goethe: Die neue Melusine
67 ring finger musik eltern palast schöne niemals kästchen stehen gewaltigen
96 kästchen gesellschaft wagen zimmer schöne beutel geld wodurch leben vergnügen

E.T.A. Hoffmann: Das Fräulein von Scudery
86 fräulein sprach schmuck meister könig arbeit that kästchen geheimniß paris

Kleist: Die Verlobung in St. Domingo
72 fremde neger fremden alte familie knaben herrn weißen bette babekan

Schmid: Mohrenfranzel
89 königin spielen burschen stück saba bursche schiffer schwarze bühne zunftmeister

Roquette: Die Schlangenkönigin
48 kahn wasser bild schlangenkönigin dorfe schlangen ufer entgegen gegend leipe

Falcon Topics also always offer thematic indicators, but as a rule they are weaker and less coherent to the extent that they are interrupted by names, proper names, and place names . In this way, they refer more to the text in question and its plot as they do (purely) to a theme. Falcon Topics, we surmise, therefore simply fall out of consideration in the case of Topic Modeling as a purely distant reading. An approach that is decidedly interested in the individual text is required in the corpus analysis in order to find it interesting in the first place. Matthew Jockers has coined a special term for Topics which are heavy with names and proper names: he refers to "topical topics,"[22] which he conceives of, however, as mere distortions of the hoped for results in the shape of thematic Topics. Our Falcon Topics are topical in their doubled function of localization. If we consider statistically calculated Topics in a hermeneutic manner based on knowledge of the text, it is above all the names, proper names, and references to places that help us to find the corresponding text even without a full-text search in the entire corpus. At the same time, Falcon Topics locate the action of the novellas in question, their setting and plot.

[22]Matthew Jockers: Text Analysis with R for Students of Literature. Cham 2014, p. 152.

If we take the idea of topical Topics seriously, it might lead us to the center of current methodological discussions surrounding Topic Modeling. We are fortunate that our *Novella Treasure Trove* corpus is part of a research group from DARIAH concerned with Topic Modeling. With the teams of Gerhard Lauer and Simone Winko (Göttingen), Fotis Jannidis (Würzburg), and Peer Trilcke (Potsdam), we are constantly and critically discussing the question of how appropriate Topic Modeling really is for literature. The group with Christof Schöch and Steffen Pielström from Würzburg that is working on improving the algorithm, is intensively thinking about why Topic Modeling analyses regularly work smoothly in the social and political sciences, whereas just as frequently problems emerge during the analysis of literary corpora. An interesting hypothesis is that the semantic fields of literary texts, in contrast to, say, parliamentary speeches or social media communication, is much more strongly layered by the setting of the texts, by words that are characteristic for the plot but not absolutely thematic. Our provisional position on this in the wake of our experiments: This layering comes to light in topical Topics. What we have described as Falcon Topics on the basis of the historical context of our novella treasure trove, can be equally understood as the result of insufficient parameter optimization and as the statistical trace of literariness.

Appendix

1	mäuse stephan hölzerne ratten höfen habicht ähnlichen topf paulus wahrhaftigkeit
2	weg nacht luft himmel sonne standen hinein schritte erde kamen
3	bertram süden spitze gipfel land osten stiefel gebüsch gänge inseln
4	pfarrer schnee hofrichter meßner förster köchin kirche warmen kellermeister stunden
5	blumen erde blume goldenen blätter gesegnet natur brunnen früchte zweige
6	nacht tag schlief bette bett uhr schlaf saß schlafen erwachte
7	schiff ufer see schiffer boot schiffe wasser wellen segel hamburg
8	bauer gott großmutter alte magd kind heiligen knabe worte seele
9	bäcker pastor baumann moor löffel schwieg busch kirche fritzens stube
10	silberburg rothenburg stadt scharfrichter wolf heyliger thal römerhöhe garten rath
11	frau graf kind grafen gräfin bruder dame verwalter frug schlosse
12	müllerin heirathen jurancon verwandtschaft base müller sich's stiefvater verlobung schwager
13	onkel großmutter julie oberförster brink trauung liebes berger brigitten vollends
14	zeit ließ schien hause alten tage wußte blieb fand that
15	gott ehre grab alte schwert herzog gnade begraben schönen letzten
16	hunde beute bloß koppigen junker herren desto dachte wild alten
17	wirthin baron gegend thurn freifrau braut lich fräulein lothar schlosse
18	wirth stiefeln netze kalender pantoffeln zusammenhalten dreifach schlingen verstecken kann's
19	rose wein sprach keller judas kellermeister apostel namen jungfer römer
20	offizier graf nacht pferde offiziere pferd husar junge husarenoffizier hauptquartier
21	müller mühle schmied justizrath actuar gerichtsdiener knabe meister schmiede müllers
22	herzog reise advocat durchlaucht baron studenten polen kurland bursch hochzeit
23	brunken maler schulmeister doctor ungeheuer kunst assessor bild freundes aufgewachsen
24	bäuerin särge dunklen blitzstrahl blitze gericht vorübergehen bärbele knopf damm
25	herr mann fort sagen frau lassen alte sehen fragte weiß
26	london grauen woche deutschen bettler england ohnedies selbstpeinigung ländlichen künftigen
27	marquis schloß paris herrn anstalt drachen zwerglein schlosse geschichte gott
28	marquis doctor professor alten jungen rom ghetto corso italien valencia
29	gewissenhaft agram damaligen ertönen heiligkeit violine romantischen ueber hauptstadt mühsam
30	landrichter erwiderte haus lappen kaufmann fjord missionär probst malanger henrik
31	küster hedeper frau küsterei wittwe meierin küsters herrn jahr schwägerin
32	leben liebe herz welt menschen seele glück herzen freude armen

33	tante notar spendvögtin glücklich freundin bloß julie marei jungfer städtchen
34	schreiber base schreibers ofen registranten keller syndicus züge erschöpfung untersten
35	dorf kessel milch schönsten dirnen schultheiß schultheißen pfanne brunnen streit
36	muhme muhme-lieutnanten striethast hanepich lene lebte erbin muhme-lieutnant-saloppel schwestern professor
37	vater kind jahre tochter haus jungen jahren kinder mädchen sohn
38	versetzte arzt wasser gräfin köln fremden ring schatulle dame felsen
39	könig königin herzog frau weib könige königs frug land volke
40	mädchen signora urballa burschen signor heirathen villa padrona signoria baldo
41	droben schwester rosine tante bruder meran bursch zehnuhrmesser hochwürden mal
42	wesen schönen gesellschaft gefühl verstand niemals gewissen schönheit leidenschaft natur
43	brief las schrieb lesen briefe schreiben geschrieben buch papier gelesen
44	vater hab laß weißt ist's komm geh kind ich's lieb
45	novelle geschichte namen form roman dichter stoff meister erzählungen krieges
46	prinzessin wiederum hauses magd rom größe römischen prinz damen gatten
47	alte bild maler kunst künstler freund bilder gemälde fremde junger
48	kahn wasser bild schlangenkönigin dorfe schlangen ufer entgegen gegend leipe
49	schnee tante mühle kindes doktor nordwind blankenheim margrets tropfen wölfe
50	schatten gold herrn grauen sonne tasche goldes seckel pracht vieles
51	kranken kranke krankheit arzt klagte heilung gehoben uebel ausdauer stock
52	ruft büchse sieht gesellen wolf bär juden fällt thiere schuß
53	mynheer meister zorgenhof amsterdam freimeister buchhalter see jüngling insel familie
54	kennen natur lernen nennen bildung kunst geist meisten höheren erscheint
55	alte schneider mädchen versetzte sohn pfarrerin gesicht burschen entgegnete bursche
56	paris deutschen frankreich töchter damen schwestern monseigneur allmählich bordeaux hamburg
57	sprach eremit freunde überall traum richter fest riefen schatz engel
58	hand augen rief stand gesicht hielt kopf trat lag mädchen
59	amtmann baron fräulein versetzte halden dame amtmanns referendar weis barons
60	heiligen nonne bild kloster antlitz kunst kirche malen bildes entgegnete
61	geld gulden thaler handel summe wasser schulden littauer wirth verkauft
62	erfreute korb betraten verleiten feine bescheidenheit fressen zahl geschirr verschiedene
63	frau conrectorin vetter vetterchen general eigentlich tone lachen erzählen weile
64	herr hahn herrn papa gast todten mama herbesheim fräulein stadt
65	treppe polizei holz ofen stufen lachen emmerich sieh niemals feuer
66	chef lieutenant freund ehre freunde gefährten ring flagge offiziere freundes
67	ring finger musik eltern palast schöne niemals kästchen stehen gewaltigen
68	jäger bauern mühle förster bauer wald thal hütte jüngling felsen

69	scheint neffe entgegnete würdest verbreiten ansicht eva ludwig paradiese streben
70	frei schuldig staatsanwalt rothmann geschworenen verhandlung schwiegersohn angeklagten männer vorsitzende
71	sprach rhein trug nebi fischer klause klotz manch berg kniggemann
72	fremde neger fremden alte familie knaben herrn weißen bette babekan
73	klarinett schloß dame wald bursch garten student fenster draußen droben
74	acker kinder bauern puppe nase steine jahren wilde aecker loch
75	wein tisch glas trinken tische essen trank gäste flasche tafel
76	major haide pferde ritten herum uwar maroshely gitter ritt lande
77	salonichi diener juden paradise griechen brittischen arbeiten athen konstantinopel alexander
78	teufel general soldaten könig hauptmann commandant offizier dienst oberst wache
79	weißen trug schwarzen haar langen schön schönen schwarze rothen blauen
80	pfarrer burg berg tubus pfarrerin gesellschaft pfarrers gleichfalls nämlich examen
81	kirche pfarrer sonntag heilige messe gemeinde priester geistliche heiligen predigt
82	theils laden wobei korb kunden fände georg's landenberger kanzlei wittwe
83	lied musik singen spielen sang talent spielte geige klavier sängerin
84	todt tod blut hülfe leiche wunde wuth todten mörder kampf
85	used total openjdk-i usr/lib/jvm/java dff acf thread xae heap jre/lib/i
86	fräulein sprach schmuck meister könig arbeit that kästchen geheimniß paris
87	doctor bauer drachen pirna erbschaft kappbauer doctors flurschütz stube pastor
88	wagen pferde fahren pferd kutscher straße reiter fremde pferden knecht
89	königin spielen burschen stück saba bursche schiffer schwarze bühne zunftmeister
90	gutsherr ohm dorf knaben förster dennoch holz uhr juden knabe
91	vermittelung gemeinde klöster gesunden gottlosen zustand dünnen gebilde gesunde eingebüßt
92	richter stadtschreiber galgen gericht urtheil herren rath strafe freiheit zuletzt
93	fenster zimmer haus thür thüre hinaus nacht treppe öffnete licht
94	vater graf mann frau marquis sehen dachte frug lies baron
95	munde vetter frau that schäferle buchenberg haus stube kapitel schäfer
96	kästchen gesellschaft wagen zimmer schöne beutel geld wodurch leben vergnügen
97	weib sag schuster lustig pfand lachen pfeife bauern frauen liebe
98	franzosen krieg frankreich könig männer vaterland volk kaiser feind französischen
99	oberst meister fräulein erwiderte leise stadt gnädiger obersten tochter gefangene
100	tanz tanzen tanze vicomte tanzte jungen musik mädchen musikanten altenkreuz