

The Order of Things. A Study on Topic Modelling of Literary Texts

Inna Uglanova, Evelyn Gius

Technical University of Darmstadt, Institute of Linguistics and Literary Studies, Dolivostraße 15, 64293 Darmstadt

Abstract

Topic modelling is considered a statistical tool for the thematic decomposition of texts. In reality, it captures only statistical patterns in the structure of the object. This sensitivity of the method for structure makes it less effective when applied to literary texts in which structure itself is a relevant feature with an artistic function. In this paper, we calculate a series of topic models for three corpora of literary narratives with various stages of data cleaning. We apply coherence values, qualitative interpretation and measurement of topic distances in order to shed some light on the regularities between text features and the quality of the topic modelling performed for literary prose.

Keywords

topic modelling, LDA, topic coherence, topic evaluation, literary texts

1. Introduction

Topic modelling is often presented by its numerous apologists as a miracle tool that magically transforms a bag of words into a series of self-describing mini stories.¹

In the field of literary studies, there are also quite successful topic modelling analyses [11], [19], [28], that repeatedly attract new victims to the literary “LDA buffet” [10]. However, the attracted people often have frustrating experiences with the method. The magic of topics seems to fail in front of literary texts. A typical output of the modelling of this type of text usually consists mainly of cohesion elements that are meaningless in the thematic sense.

Cohesion elements are the basic building blocks of the textual surface. They form the basic framework of a text, namely the structure of the dependencies between the text elements.²

Consider, for example, the following topic from one of our data sets:


Topic 3 [Subcorpus “Gutenberg narrative”, whole texts]: mein – mich – haben – unser –

CHR 2020: Workshop on Computational Humanities Research, November 18–20, 2020, Amsterdam, The Netherlands

✉ inna.uglanova@tu-darmstadt.de (I. Uglanova); evelyn.gius@tu-darmstadt.de (E. Gius)

🆔 0000-0002-8092-3512 (I. Uglanova); 0000-0001-8888-8419 (E. Gius)

© 2020 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

¹See for example [2, p. 1009]:

Topic “ARTS”: new - film - show - music - movie - play - musical - best - actor - first -
york - opera - theater - actress - love

²Coherence, in contrast, forms the functional-communicative core of the text [18]. Coherence elements constitute the content, the meaning of a text. Therefore, whereas cohesion performs a language-forming function in a text, coherence has a text-constituting function, as for example in the topic “ARTS” by [2] cited above. In a literary text, coherence elements form a holistic symbol-forming unit.

nicht – nach – jahr – durch – abend – über – noch – kein – freund – kommen – stunde.³

A reasonably creative interpreter may certainly find a meaning in this word list. Without context, however, it is too open for a thematic assignment and therefore it is difficult to come to an interpretation that interprets the word list as a topic and has a certain degree of plausibility. Such uninterpretable topics are common in a first attempt to topic modelling in a literary corpus. Topic modelling in the context of literary studies thus seems to miss something fundamental for literary texts. There seems to be a systematic problem that causes this frustration.

On the other hand, from the experience of topic modelling literary texts, we also know that it works very well if the texts to be modelled are optimised accordingly: They must be clean, segmented, etc. What is the reason for this? We will try to find an answer to this question by experimenting with different constellations of data preparation and cleaning and relating the outcomes to computational and manual evaluation.

2. Research Design

2.1. Main Hypothesis and Objectives

In this paper we try to shed some light on the mechanisms behind the functioning (or non-functioning) of topic modelling for literary texts. The successful practices of topic modelling of literary texts discussed above suggest that especially the manipulation of text structure has an impact on the results of topic modelling. This was the main hypothesis behind the approach we present here. We assume that the effects of these manipulations should be recognizable both quantitatively and qualitatively. This should be possible for different data sets if they are similarly parameterized with regard to the manipulations.

In order to verify this hypothesis, we have performed the following:

- We systematically tested different possibilities to eliminate or minimize redundancy and their influence on topic modelling, and
- assessed the outcomes quantitatively and qualitatively.

More precisely, we have evaluated the various manipulations of the texts by calculating coherence values in order to be able to make general observations (cf. section 3). Thereafter we took a closer look to some of the test models, focusing on the best, or at least a good model for each data type. By using both quantitative and qualitative approaches we tried to understand how the manipulation of data influences the quality of the topics we can model (cf. section 4). Based on this we finally were able to make some observations on topic modelling for literary texts (cf. section 5).

2.2. Data Configuration

This study is based on the hermA corpus [7] which has been compiled from KoLiMo [8]. KoLiMo contains texts from the three major text repositories Textgrid [25], Deutsches Textarchiv [5] and project Gutenberg [16]. The hermA corpus is a collection of the about

³translation: my - me - have - our - not - after - year - through - evening - over - yet - no - friend - come - hour.

1,800 German-language prose texts from the period between 1870 and 1920. The texts have been lemmatized, tagged and provided with metadata.

In order to model the behavior of the text structures, two further subcorpora were created from the hermA corpus: “illness” (44 texts) and “Gutenberg narrative” (354 texts). These subcorpora are more homogeneous regarding content or structure of the prose texts.

The subcorpus “illness” is composed of narratives where the illness of one or more protagonists is a relevant issue. Therefore, the subcorpus “illness” is to a certain extent homogeneous with regard to its content. This could facilitate thematic modelling or the recognition of the thematic structure by the tool.

The second subcorpus consists of texts originally taken from the Gutenberg project that are explicitly marked as narrative. Although the genre classification in the Gutenberg Project is not particularly reliable, we consider “narrative” a label that is a little controversial. Due to its more general nature it also is very unlikely to contain false positives, i.e. prose texts, that should not be considered narratives. Due to the classification of these texts as the same genre we assume that they have a comparatively homogeneous text organization. Since topic modelling is sensitive to structure, this could improve the outcome of the tool.

For preprocessing of the data, typical operational steps for topic modelling were carried out:

- removal of numbers, special characters, punctuation
- removal of words less than or equal to three characters long (these normally have no effect on the thematic profile of a text)
- conversion of the lemmas to lower case

For the analyses, partial segmentation (text chunks of 300 or 500 words) and the restriction to certain word-formation classes (NN/normal nouns, ADJA/attributive adjectives, ADJD/adverbial or predictive adjectives, VVFIN/finite full verbs) were carried out as optional additional text manipulation. The choice of segment length was determined by the established practice in Computational Linguistics and Computational Literary Studies (cf. [28], [20]) as well as by the properties of the data set. Due to the fact that our corpus consists of very heterogeneous texts, the segment lengths should be neither long nor short.

The choice of the word-formation classes was determined by their belonging to the so-called “full” words (autosemantics). In contrast to synsemantics (“functional” words like articles, prepositions, pronouns, etc.), these are the words that determine the thematic profile of the analysed texts.

For all three corpora, we tested three data types that implement these aspects:

1. whole texts
2. segmented texts
3. word-class lemmas (NN, ADJA, ADJD, VVFIN)

Each data type was tested with and without further cleaning. The data cleaning consisted of removing the high-frequency and low-frequency words (hereafter: frequencies) from the lemma lists as well as stop word removal.

Words were considered low-frequency words when appearing in a maximum of three texts of the corpus. The threshold for high-frequency lemmas was set at 0.5. This means that those words that occur in more than 50% of all texts were removed. These thresholds were tested in order to grasp the words that are too specific (and therefore do not reflect a topic of a text)

as well as words that are too general (and therefore little informative). The filtered units are considered excessive noise and their removal should make thematic modelling easier.

Stop words have been eliminated with a stop word list consisting of 1,111 units. It was created from two different sources. One part (620 units) contains mainly the most common functional words of the German language [6]. The second part (491 units) was taken from the subcorpus “illness”. These units were highly frequent words marked as named entities in the lemmatized corpus.

An overview of the tested variants and respective outcomes is given in Table 4.

2.3. Modelling and Evaluation Methods

The topic modelling was carried out using the most popular technique from the class of probabilistic generative models, Latent Dirichlet Allocation, developed by [2]. The analysis was performed by the topic modelling module implemented in a MALLET (MACHINE Learning for Language Toolkit) package [13]. For statistical inference, this software uses the Gibbs sampling algorithm.

For the number of iterations, we used 1,000, the value set as default in MALLET. The value of the optimization interval has been set to 5 because the interval has proven to be more suitable than the default of 10 in tests.

The quality of the calculated models was evaluated using the coherence measure Cv which systematically yields the best results and has become the state-of-the-art method in topic modelling [17] (cf. section 3).

We also used Cv for identifying the best model for a specific data set (cf. subsection 4.1). For these models, an additional qualitative evaluation was carried out for all relevant data sets (cf. subsection 4.2).

Finally, we used the visualization provided by [23] for a global view of the topic models based on the distance between topics in order to have an additional perspective on the impact of data cleaning on the quality of topic modelling (cf. subsection 4.3).

The modelling of the subcorpora was performed on a standard computer. The modelling of the hermA corpus as well as all calculations of the coherence value and tests of various model parameters were performed on the Lichtenberg high-performance computer at the Technical University of Darmstadt.

3. Coherence Values

3.1. Calculating Cv

The coherence measure Cv is based on a stepwise increase of the number of topics using a sliding window algorithm. A coherence value is calculated for each step (window). The essence of the method is to measure the contextual (statistical) dependencies between words within a topic. The method is based on a combination of two methods: normalized pointwise mutual information and cosine similarity.⁴ By varying the window, the best topic parameterization is determined. The coherence values lie in the interval $0 < w < 1$. The higher the value obtained, the greater the coherence between the words, the better a topic or model should be.

⁴For a detailed and understandable explanation cf. [24].

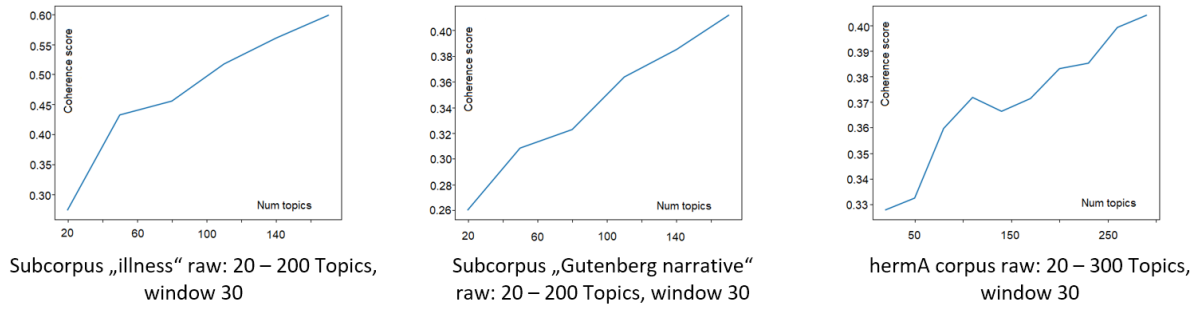


Figure 1: Dependence of coherence on the number of topics (whole texts)

3.2. Coherence Values for Whole Texts

3.2.1. Whole Texts without further Cleaning

For whole texts without further cleaning, the change rate of coherence values is, with some fluctuations, directly proportional to the number of topics (cf. Figure 1). In other words, the larger the number of topics, the greater the coherence value. The greater this number is, the more differentiated the obtained thematic structure of the analysed data should theoretically be. However, with the increasing number of topics, there is a risk that the topics will become too detailed and the thematic structure will break up into small, difficult to interpret pieces. The obtained values confirm this presumption. The topics do not seem to reflect the general thematic structure of the texts, but rather, as we have found in the qualitative analysis, the thematic structure of individual texts (cf. subsection 4.3). However, this is not the aim of topic modelling.

Another problem of increasing coherence values is that it is impossible to find the best parameterization for a model on the basis of continuously increasing values which is the underlying objective of the coherence measure Cv .

3.2.2. Cleaned Whole Texts

After the removal of frequencies and stop words, the coherence values of the two large data sets (subcorpus "Gutenberg narrative" and hermA corpus) have changed distinctly. This can be seen from the shape of the curves which change radically their direction.

As can be seen in Figure 2, the coherence values are inversely proportional to the number of topics: The greater the number of topics, the smaller the coherence values. Another important difference compared to the whole texts without cleaning is that the range of coherence values changes. The coherence values become much higher than for the whole texts, i.e. the quality of the topics improves clearly.

After the removal of high-frequency and low-frequency lemmas the development of the coherence values is discontinuous. The additional elimination of stop words, however, brings clear improvements regarding the continuity of the coherence values. The irregularities we see could be the result of a transition between two levels of information (i.e., from full text to cleaned text without redundancies). This observation requires further verification. Presumably, the development becomes homogeneous again when the language redundancy (cohesion elements) is eliminated and the texts become more consistent in a statistical sense.

Only for the small data set, the subcorpus "illness", no improvements are observed after

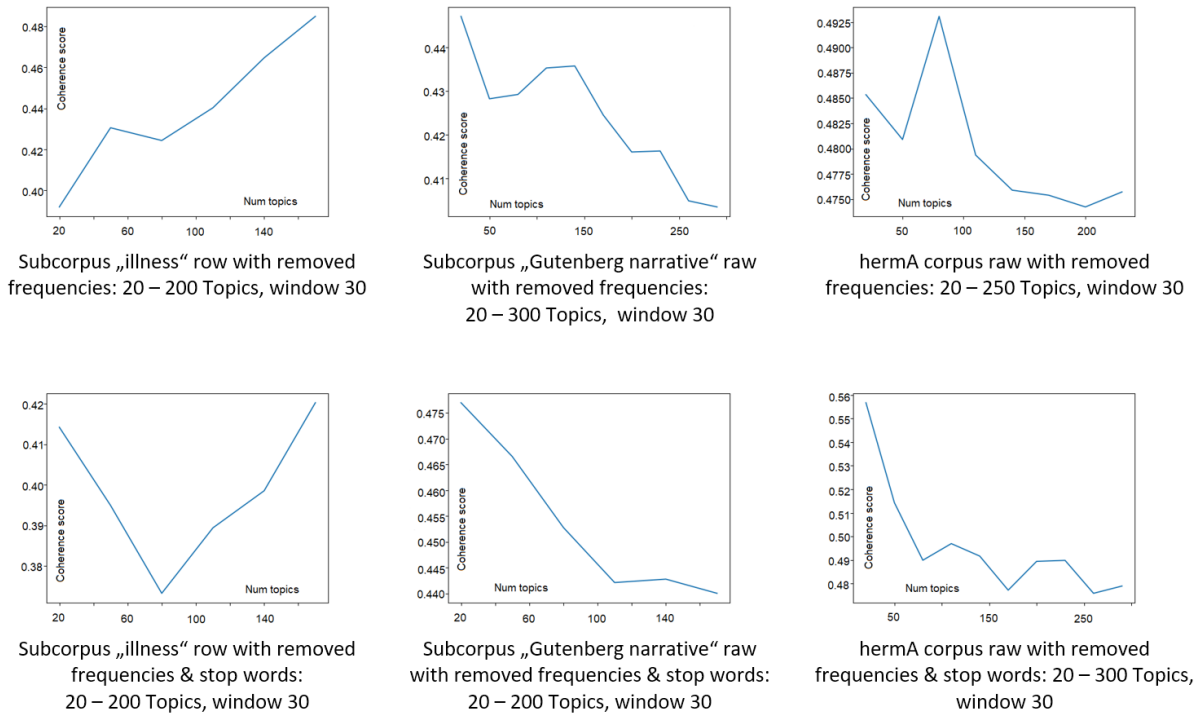


Figure 2: Dependence of coherence on the number of topics (cleaned whole texts)

the cleaning. Although the coherence values have improved too (cf. coherence values for test models in Figure 4), the general trend remains the same as for the raw data. The distribution in the subcorpus reacts to the cleaning, which is evident from the change in the shape of curves. However, for a tool like topic modelling which works with mass phenomena, there is probably not enough data for statistical inference in our subcorpus “illness”.

3.3. Coherence Values for Segmented Text

3.3.1. Segmented Texts without further Cleaning

The segmentation of the texts leads to some changes in the relation between the coherence values and the number of topics: The curve rises like a parabola (cf. Figure 3). Based on the shape of the curve, it can be assumed that a certain balance will slowly develop between the two parameters. This means that increasing the number of topics will not bring any significant improvement in their quality because the coherence values will slowly stabilize.

The observed effect is a result of the segmentation of the texts. Since the length of the text segments is fixed, with the number of topics increasing, the structure of the topics becomes gradually more homogeneous. All parts of the topic structure become equally relevant and further iterations do not result in new possible combinations. In physics, such phenomena are referred to as the entropy of the system having reached its maximum.

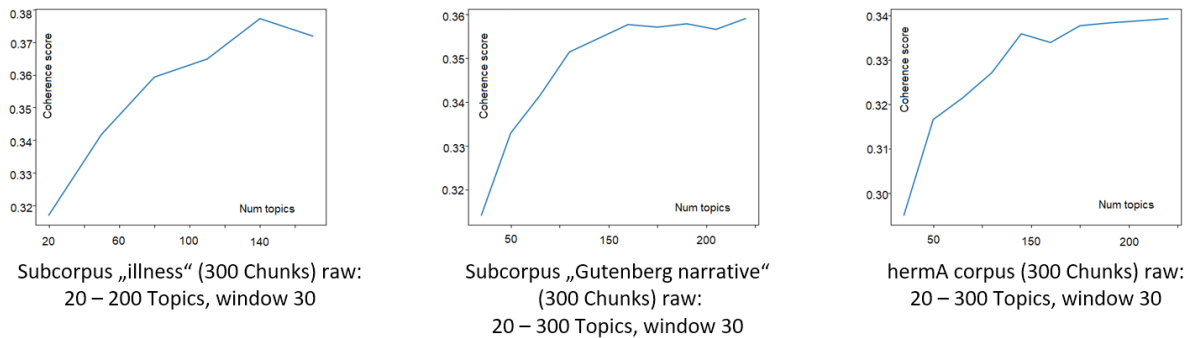


Figure 3: Dependence of coherence on the number of topics (segmented data without cleaning; selection)

3.3.2. Segmented and Cleaned Texts

Interestingly, the cleaning of the segmented data has no significant impact on the curve shape of the coherence values (cf. Figure 4). Although you can see that the data structure reacts to the elimination of the text elements, the curve still retains its basic shape. However, the quality of the coherence values is getting better with every cleaning stage. This means that the size of the data set in combination with the removal of frequencies and stop words has a positive effect on the topic quality.

Why does the curve not change its shape in the same way as it does for whole texts? This is probably one of the consequences of the segmentation. In the segmented texts fewer lemmas have been removed in comparison to the non-segmented texts (cf. column “Number of tokens” in Table 4). Since frequencies are calculated based on text segments (and not on the whole texts), the frequency structure changes less strongly here than when frequencies are calculated for non-segmented texts.

3.4. Coherence Values for Selected Word Classes (all Variants)

The restriction of the lemmas used to certain word classes is based on the idea that it makes the text structure more homogeneous and thus more suitable for topic modelling. This assumption is also confirmed by the coherence values. The curve shapes follow an already known pattern that was observed by the whole texts. This holds both for the raw and the cleaned data sets (cf. Figure 5 for the graphs for the subcorpus “Gutenberg narrative” and the hermA corpus). However, their character has changed qualitatively: The dependence between the coherence values and the number of topics becomes closer and stronger. An almost linear dependence between the parameters is observed which, for now, confirms the assumption about the homogeneity of the thematic structure.

The subcorpus “illness” is again an exception. The raw texts of this data set follow the general pattern, whereas the removal of frequencies and stop words results in an irregular distribution. For example, it is impossible to find any trend in the data with removed frequencies (cf. Table 1). In order to interpret the observed results correctly, the content structure of the respective topics must be taken into account. The character of the graphs leads to the assumption that after the text cleaning, the different semantic systems collide with each other and two topics coincide in one topic.

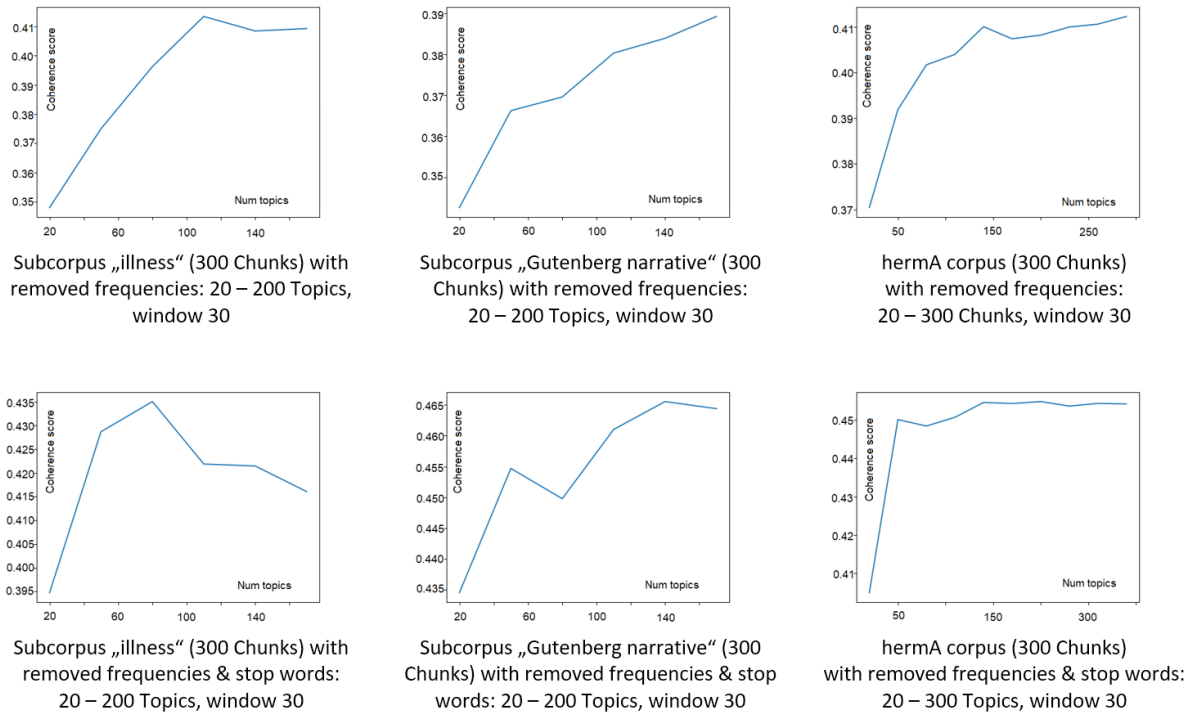


Figure 4: Dependence of coherence on the number of topics (segmented data with cleaning; selection)

Table 1

The dependence of coherence on the number of topics (subcorpus "illness", word classes, without frequencies)

Number of Topics	Coherence values
20	0.5297
50	0.5322
80	0.5258
110	0.5194
140	0.5321
170	0.5380

4. Evaluation of the Test Models

4.1. Coherence Profiles

4.1.1. Observed Coherence Types

When taking a closer look at the coherence values, we can derive four coherence types which correspond to the respective configuration type of the data sets:

1. **ascending coherence values** correspond to non-segmented raw data
2. **descending coherence values** correspond to non-segmented cleaned data
3. **parabolic coherence values** correspond to segmented data, both raw and cleaned
4. **discontinuous coherence values** correspond to non-segmented data and data with selected word classes in both cases from the corpus "illness" with cleaning, the only data set reacting unstably to the structural change

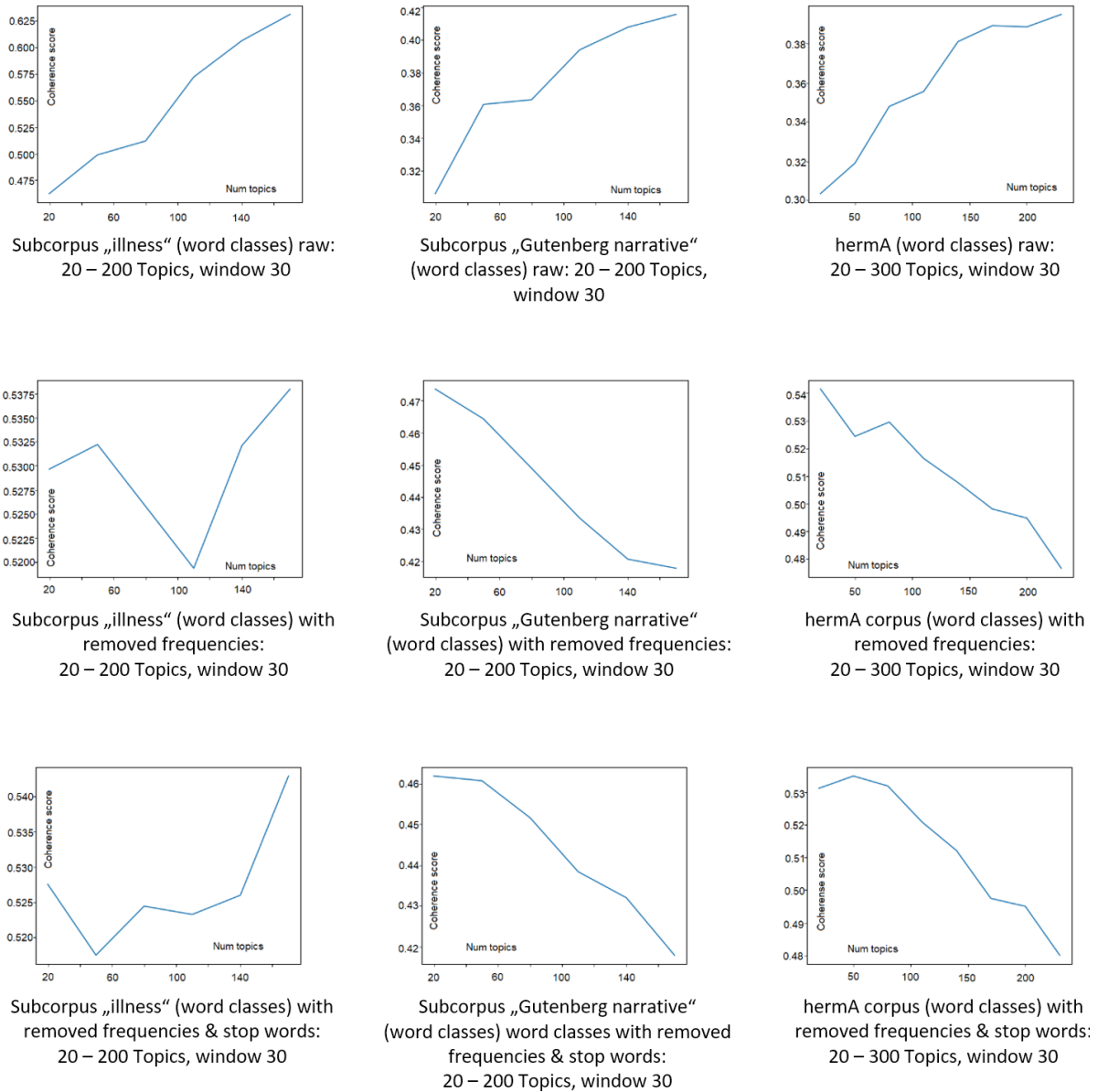


Figure 5: Dependence of coherence on the number of topics (word classes)

It is particularly interesting that the highest coherence value does not seem to necessarily correspond to the quality of topics. As already discussed, it is improbable that the increasing tendency of the first type correlates de facto with the quality of topics in the thematic sense. One could assume that the observed slope is caused by the coherence between the cohesion structures which, for example, connect proper names or modal verbs to topics. However, the final interpretation can only be made after the qualitative analysis of the topic structure has been performed. It should show which qualitative properties are associated with the observed coherence value trends (cf. subsection 4.3).

Table 2
Best Coherence Values

C_v	Data type (& Corpus)
0.5018	word classes raw & stop words (hermA)
0.5276	word classes raw ("illness")
0.5338	word classes & stop words ("illness")

4.1.2. Identifying the Best Topic Models

The best configuration “Coherence – Number of Topics” is the one where the number of topics corresponds to the best coherence value. For ascending, descending and variable types of changes, the best configuration can therefore not be defined straightaway. For our data it was possible only in a few cases to determine the best topic model on the basis of the calculated coherence values (cf. Table 4).

For the determination of the number of topics to be calculated, we assumed as a rule of thumb that the number of topics should lie between 100 and 150. With this specification, the topic structure is neither too detailed nor too general which makes this interval the ideal measure for modelling literary texts.⁵ The concrete number of topics used for further investigation was then determined on a case-by-case basis since the data sets differ in quality (structure) and quantity (size). In general, we attempted to remain consistent within a data type in terms of the number of topics and to take into account the general curve progression of the observed coherence types.

The range of obtained coherence values lies between 0.3184 (hermA corpus, segmented raw data) and 0.5338 (subcorpus “illness”, word classes without stop words). The highest values were found in the group using only selected word classes (cf. Table 2). Since the general distribution of the data is already known (cf. Figure 5), it can be assumed that, apart from the value of the hermA corpus, the two other results are probably less informative.

Most of the values for the data with cleaning are in the interval [0.43; 0.5] which indicates a low coherence between the words in the topics.

No consistent picture can be drawn from segmented and non-segmented raw texts. The highest values were also found for the subcorpus “illness”. The segmentation of the texts in the hermA corpus even has a negative effect on coherence (0.3647 for the whole texts against 0.3184 for the segmented ones). In the subcorpus “Gutenberg narrative” this proportion is slightly shifted in a positive direction in favour of segmentation (0.3409 for the non-segmented raw texts against 0.3578 for the respective segmented ones).

In general, it can be said that the coherence values are not absolute and therefore not comparable. They vary greatly from one data type to another. The values that are good for one datatype can be bad for another datatype which makes them not comparable. However, this does not mean that coherence values are bad. This parameter describes the coherence structure of only one concrete data type within which it can be compared, for example only within the segmented data (in 300 words) of the hermA corpus. Nevertheless, the following applies to all coherence values: As the data sets are cleaned, an increase in coherence values is observed. This means that coherence can be considered a function of the cleaning operation.

⁵For the hermA corpus, however, 200 topics were tested once to see whether this would lead to any significant changes in the quality of the topics (cf. in Table 4 the data set with removed frequencies and stop words).

In order to further evaluate the observed data, the qualitative analysis of the topic structure was carried out.

4.2. Qualitative Evaluation of the Topics

4.2.1. “Good” Topics

Even though coherence values can be very helpful for the determination of good models, from a humanist perspective human evaluation is still the gold standard in the evaluation of topic models [4]. Therefore, we performed a manual evaluation of the interpretability of the topics of the selected test models.

Generally, the first ten words in a topic determine about 30% of its content. That is why ten words per topic are set as a default value in the most commonly used algorithms (cf. [14]). The higher the interpretability threshold is, the greater is the variation in subjective estimations. Decreasing the default threshold reduces the individual variability in assessments and thus makes them stable.

In the present study, a topic is considered to be well interpretable if at least 5 words are related to a common thematic aspect. Consider for example the following topic:

Topic 83 [Subcorpus “illness” raw, segmented (300 words)]: werden – durch – nach – **richter**
– **gericht** – fischer – können – ursinus – **untersuchung** – selbst – **mörder** – **verbrechen**
– sondern – **erklären** – dieser.⁶

This topic has six words (judge - court - investigation - murderer - crime - explain) among the ten most relevant words that can be connected to court.

In order to identify the ratio of “good” topics, we inspected all 3,430 topics of the test models and classified them interpretable or not, depending on the presence of at least five words with a shared thematic structure.⁷

As can be seen from Table 4, as the percentage of good topics increases, the coherence value also increases. Therefore, we measured correlation between three main coherence types (ascending, descending, parabolic) and the values of interpretability. For this we used Kendall’s tau-b correlation that can deal with the data with many tied ranks. The value of the coefficient is 0.531 which means that there is a moderate correlation between these two parameters, i.e., the coherence value and the ratio of “good” topics.

4.2.2. Thematic Structure

What does topic modelling tell us about the thematic structure of the corpus in general? What stories have been told between 1870 and 1920? What topics were of interest for society at that time?

The core of the thematic diversity of the given data consists of about 100 themes or thematic complexes. They are distributed in different proportions and qualities through all data sets. As can be seen from Table 3, the three most common themes belong to the thematic complexes “bourgeoisie / society”, “nature” and “colonial history”.

⁶translation: are - by - after - **judge** - **court** - fisherman - can - ursinus - **investigation** - himself - **murderer** - **crime** - but - **explain** - this.

⁷This qualitative evaluation was performed by only one person. Even though she is an expert with extensive experience in analysing language data, we would recommend to enhance such evaluation approaches to at least two evaluators for a higher degree of reliability. Since qualitative evaluation was not the focus of our study and the number of topics was very high, we could not implement it here, yet.

Table 3
Most frequent thematic topics

Thematic complexes and topic occurrence			
Society & Bourgeoisie	165	History	35
Nature	105	Seafaring	35
Colonial History	55	Feelings	34
Family	49	Empire	32
Church & Belief	47	Arts	29
Nobility	41	Justice & Crime	28
Body	39	Military	26

In literary texts, topic modelling brings to light the basic thematic structure of human existence and experience – the order of things: What humans are interested in (arts, theatre, music, history, adventure, state life, world politics, etc.); what humans engage in (hunting, dancing, building, trading, crime, etc.); what inspires humans (aviation, poetry, reading, writing, friendship, family, etc.) and what makes a human being sad (fears, illness, grief, death, etc.). It reproduces the social environment (empire, bourgeoisie, state life, war, etc.) and the typical patterns of everyday life (eating, drinking, going out, entertaining, clothing, studying, communicating, belief, etc.). All this is packed into topics.

4.3. Topic Distribution

For further evaluation of the models we used visualizations created with pyLDAvis [23] in order to compare the relevance and the distribution of topics in the various data sets. The topics are plotted as circles whose diameter shows their prevalence in the corpus. Their position in the plane is determined by the computed distance between topics.⁸

As we will see, the cleaning of data affects the ratio between cohesion and coherence (by reducing cohesion). This, in turn, leads to qualitative and quantitative transformations affecting the semantic structure of the data as can be observed in the visualization.

The difference in the thematic structures between the first and second types of coherence, i.e. ascending and descending coherence values, becomes even clearer when visualizing selected constellations.

The first type is dominated by cohesion structures (large blue circles in Figure 6). They are attractors which neutralize as dominant topics the actual “thematic” structure. The typical topics look like this one:

Topic 2 [hermA corpus, raw, whole texts]: sein – sich – welcher – werden – haben – nicht – dies – können – aber – dieser – nach – auch – derselbe – jetzt – noch.⁹

The thematic topics, however, are mainly concentrated on the periphery of the distribution (cf. Figure 6, left graph, lower right quadrant). After the removal of stop words and frequency words, the text structure appears more differentiated thematically: The structure becomes more compact and diffuse at the same time (cf. Figure 6, right graph).

⁸For a detailed explanation cf. [23].

⁹translation: to be - oneself - which - will - have - not - this - can - but - this - after - also - same - now - still.

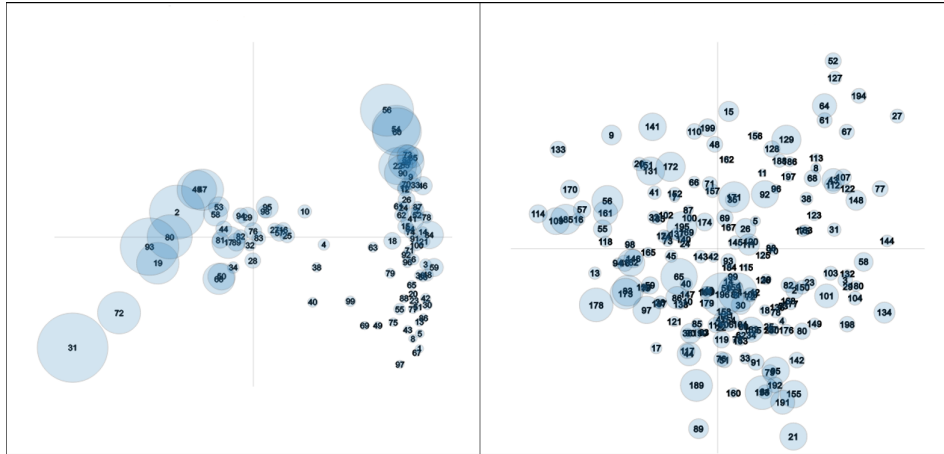


Figure 6: Topic distribution in hermA corpus: data without cleaning (left), data with removed frequencies and stop words (right)

If the topics of the first coherence type (ascending curve) are compared, an effect of the size of the data set becomes visible. The internal topic quality increases with the size of the data set, see the structure of the topic “seafaring”:

Topic 22 [Subcorpus “Gutenberg narrative” raw, whole texts]: **schiff** – eduard – wonström – **boot** – **land** – **kapitän** – jetzt – können – **wasser** – hans – welcher – nicht – hier – beide – **bord**.¹⁰

Topic 15 [hermA corpus raw, whole texts]: **schiff** - **kapitän** - **boot** - **meer** - **insel** - **wasser** – **bord** - **land** – vater - **segel** - **matrose** - **wind** - **deck** – tier - **hafen**.¹¹

The simply removing of the frequencies (coherence type 2, descending) significantly facilitates access to the thematic structure of the text. This can be seen in the following topic which belongs to the same data set as topic 22 above but has been modelled on the cleaned texts:

Topic 40 [Subcorpus “Gutenberg narrative” without frequencies]: **schiff** – **segel** – **boot** – **kapitän** – **bord** – **sturm** – **deck** – hans – **meer** – **steuermann** – **fisch** – fahrzeug – **mast** – lord – **matrose**.¹²

The segmentation (coherence type 3, parabolic) brings out the thematic structure of the data set. The dominance of the cohesion elements is weakened by their distribution across the segments. However, they now appear in the proper thematic topics which makes them more ambiguous. The topic distribution confirms this observation (cf. Figure 7). Since the elements of cohesion and coherence are mixed, there is no obvious opposition between both structure types in comparison to the Figure 6. That is why the distribution does not change significantly its shape after the removal of frequencies and stop words. However, with the additional cleaning, the words in the topics seem to become more concrete in their meaning (the topics remain the same), as can be seen here:

¹⁰translation: **ship** - eduard - wonström - **boat** - **land** - **captain** - now - can - **water** - hans - which - not - here - both - **board**.

¹¹translation: **ship** - **captain** - **boat** - sea - island - water - board - land - father - sail - sailor - wind - **deck** - animal - **port**.

¹²translation: **ship** - sail - **boat** - **captain** - board - storm - **deck** - hans - sea - **helmsman** - fish - vehicle - **mast** - lord - **sailor**.

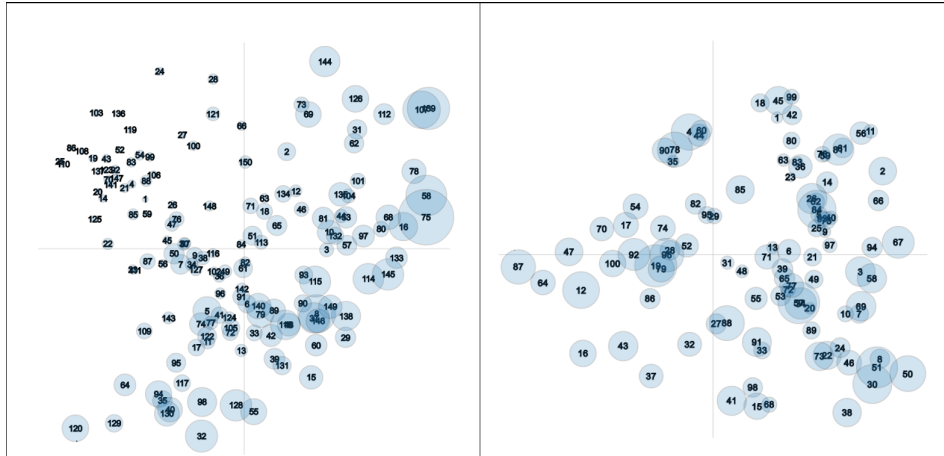


Figure 7: Topic distribution in the segmented hermA corpus: raw data (left), data with removed frequencies and stop words (right)

Topic 21 [hermA corpus, raw, whole texts]: pfarrer – kirche – kaplan – herr – bauer – heilig – johanneses – lehrer – kloster – mönch – beten – lesen – reden.¹³

Topics 157 [hermA corpus, whole texts, without frequencies & stop words]: kloster – mönch – papst – geistlich – bischof – fromm – nonnen – katholisch – priester – sünde – zelle – kanzler – weltlich – römisch – sankt.¹⁴

Topic 7 [hermA corpus, raw, segmented]: sein – gott – kirche – heilig – herr – werden – pfarrer – beten – priester – dies – fromm – aller – unser – über – kloster.¹⁵

Topic 6 [hermA corpus, segmented, without frequencies & stop words]: kirche – heilig – herr – priester – kloster – mönch – fromm – beten – bischof – jude – christ – himmel – pater – bruder – gebet.¹⁶

The visualization of the topic distribution for the data of the fourth coherence type which is characterized by its instability, confirms the quantitative results that were discussed earlier (cf. Figure 8): The removal of the cohesion elements results only in an additional fragmentation of the thematic structure. Due to the limited amount of data, the potentially thematic topics are only recognizable if they have already been seen in other data types. However, one can observe here the process of formation of the topics, cf.:

Topic 16 [Subcorpus “illness”, word classes without stop words]: knochen – strosack – madame – einbildung – allerhöchst – gebiet – fürstlich – leidig – kümmerlich – verwirrung – **mord** – **spur** – meinung – **schlagen** – morgend.¹⁷

¹³translation: priest - church - chaplain - lord - farmer - holy - johanneses - teacher - monastery - monk - pray - read - speak.

¹⁴translation: monastery - monk - pope - spiritual - bishop - pious - nuns - catholic - priest - sin - cell - chancellor - secular - roman - sanct.

¹⁵translation: to be - god - church - holy - lord - become - priest - pray - priest - this - pious - all - our - about - monastery.

¹⁶translation: church - holy - lord - priest - monastery - monk - pious - pray - bishop - jew - christ - heaven - father - brother - pray.

¹⁷translation: bone - strosack - madame - imagination - highest - area - princely - tiresome - wretched - confusion - **murder** - **trace** - opinion - **beat** - morning.

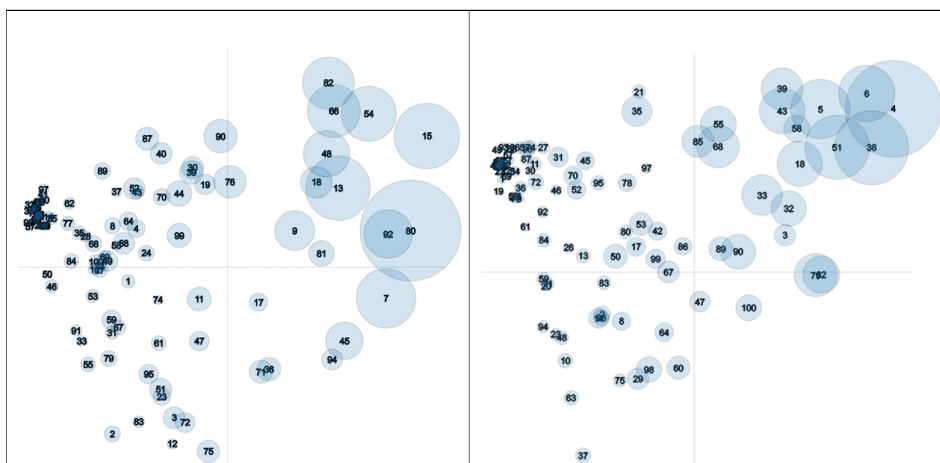


Figure 8: Topic distribution in the subcorpus “illness”: whole texts without stop words (left), word classes without stop words (right)

Topic 88 [Subcorpus “illness”, without stop words]: leben – taler – ursinus – schreiben – **untersuchung** – **richter** – wilke – berlin – rosenfeld – **gericht** – **gift** – kaltofen – miltenberg – februar – bleiben.¹⁸

These kinds of topics are a kind of proto-topics which will later be crystallized into “real” topics by further structural changes of the data set.

5. Topic Modelling for Literary Texts: Some Observations

As discussed in our introduction, the magic of topic modelling often seems to fail for literary texts. At least, if the texts are not cleaned. The experiments and evaluations we performed pointed us not only to solutions, but also to possible explanations for the problems one typically encounters.

One general problem for topic modelling is probably the phenomenon of language redundancy. This has so far only been reflected to a limited extent. Speech is 50-55% redundant [15, 22] and therefore it is not surprising that the majority of the output of topic modelling is not very informative. Frequently used words such as function words, proper names, etc. dominate the topics and their prevalence can obscure the thematic structure of the text to a certain extent. This is also the case in studies with literary texts [9, 26, 27]. Our experiments confirmed that the different approaches of cleaning the data improve the outcomes of topic modelling. Therefore, one should try to get rid of redundant units when preparing the data for modelling.

However, when applying topic modelling to scientific discourses [1] or diaries [3], good results are obtained without or only with little cleaning of data.¹⁹ For understanding these differences

¹⁸translation: life - thaler - ursinus - writing - **examination** - **judge** - wilke - berlin - rosenfeld - **court** - **poison** - kaltofen - miltenberg - february - stay.

¹⁹Cf. the conclusion of [21] who specifically tested the effect of the cleaning of data on non-fiction texts (ArXiv papers, New York Times articles, Reuters data set, biographies from IMDb, etc.): “Except for the dozen or so most frequent words, removing stopwords has no substantial effect on model likelihood, topic coherence, or classification accuracy”. Interestingly also present in the successful cases of topic modelling, there are “problematic” topics, i.e., topics that are either difficult to interpret or not very informative for us. But

one should bear in mind that topic modelling is a statistical procedure for recording regularities in a data set. It can model all semiotically interpretable systems from gene structures to images by redistributing the entities to be studied into statistically coherent groups based on their common occurrence in a data set. The term ‘topic’ is therefore not to be understood as a synonym for theme but rather as a metaphor for objects with similar structural, functional or other characteristics collected in a coherent group. Topic modelling draws no connection the content side, i.e. it does not address the meaning of the signs it deals with. Therefore, even though topic modelling is typically used for the content analysis of texts, we should keep in mind that topic modelling operates solely on forms and structures, i.e. on structural phenomena.

The difference in the results of topic modelling between scientific texts or diaries and literary texts is therefore probably related to the structure of the texts. The former are characterized by a standardized structure that renders them ideal objects for topic modeling. Literary texts, on contrary, are characterized by a special type of information organization in which the text structure itself is part of the artistic function. This means that literary texts are determined, among other things, by their structure. Moreover, this structure can have many different forms. This is more obvious for poetry,²⁰ but it also holds for prose. In literary texts, structure can have its own meaning. If the structure is changed, the meaning is changed accordingly.

Now, topic modelling brings to light the regularities of language structure. As a merely statistical tool, topic modelling shows the regularities or irregularities manifested in form. ‘Form’ here means the representation of text as a “bag-of-words”. Therefore, topic modelling works not on the structure of texts but instead on a statistical matrix of word frequencies in these texts.

Any manipulation of texts leads to a functional change in the structure which manifests itself above all in the reorganization of the qualitative and quantitative ratio between cohesion and coherence in the data set in favour of the latter. Without manipulation of the texts, the output of topic modelling is therefore dominated by predominantly cohesive text elements such as pronouns or conjunctions. With transformation (cleaning, segmentation), the structure acquires new qualities. These structural changes are reflected in the respective coherence types we have discussed. The less cohesion elements the structure contains, the more heterogeneous and complex it becomes in the thematic sense. This results in better topics.

For the deployment of topic modelling for literary analyses we have to take into account that it ignores an essential aspect of a literary text: literariness. The statistical method shows the semantic and linguistic regularities or irregularities that reproduce the order of things, their basic patterns. Therefore, topic modelling is not suitable for the evaluation of the literary quality of texts or an analysis of texts that includes their literary quality. For tackling literary quality research on the adequate processing of the texts is still needed.

But topic modelling can be used to a certain extent for a more content-related analysis of literary texts, if it is applied to a certain quantity of appropriately cleaned texts. As we have shown, the best results are achieved with the combination of segmentation and cleaning. The size of the data set also levels the dominance of the cohesion structures and yields better results.

they are normally not addressed in publications.

²⁰Cf. Lotman’s observations on structure in poems: “A change in structure gives the reader or observer a different idea. It follows that there are no ‘formal elements’ in a poem in the sense that one usually associates with this term. An artistic text is a complexly constructed sense. All its elements are elements carrying meaning” [12, p. 27].

Acknowledgments

Extensive calculations on the Lichtenberg high-performance computer of the Technical University of Darmstadt were conducted for this research. The authors would like to thank the Hessian Competence Center for High Performance Computing – funded by the Hessen State Ministry of Higher Education, Research and the Arts – for helpful advice. We would like to express our special thanks to the department staff of generic services Christian Griebel, Tim Jammer and Dr. Benjamin Juhl.

Table 4
Data overview

Corpus	Segmentation	Number of texts	Number of authors	Cleaning (frequencies/ stop words/ word classes)	Number of tokens (MALLET)	Number of text segments	Coherence trends for window 30	Number of topics	Test Model		
									Coherence values	"good" topics (%)	
	-	44	34	-	1 364 768	44	increasing	100	0.4427	2	
		44	34	-	1 364 768	44	increasing	150	0.4955	5.33	
		44	34	stop words	802 205	44	decreasing/increasing	100	0.4673	9	
		44	34	word classes	702 865	44	increasing	100	0.5338	18	
	300	-	44	34	word classes/ stop words	637 662	44	unstable/increasing	100	0.5338	18
			44	34	-	1 364 768	9 522	140 / 0.3772	100	0.4302	26
			44	34	frequency	1 037 515	9 522	110 / 0.4135	100	0.4484	37
			44	34	frequency/ stop words	724 472	9 522	80/0.435	100	0.4425	42
	500	-	44	34	-	1 364 587	5 715	increasing - flat	110	0.4223	23.64
			44	34	frequency	960 328	5 715	100 / 0.4093	100	0.4505	26
			44	34	frequency/ stop words	719 431	5 715	80 / 0.4295	110	0.4535	52.73
			354	136	-	5 339 036	354	increasing	110	0.3409	6.36
354			136	frequency	1 699 950	354	decreasing	110	0.4347	34.55	
354			136	frequency/ stop words	1 651 291	354	decreasing	110	0.4595	38.18	
-	-	354	136	word classes	2 934 529	354	increasing	100	0.3789	23.00	
		354	136	word classes/ frequency	1 384 994	354	decreasing	100	0.4335	45.00	
		354	136	word classes/ frequency/ stop words	1 366 483	354	decreasing/ decreasing	100	0.436	43.00	
		354	136	word classes/ frequency/ stop words	1 366 483	354	decreasing	50	0.445	48.00	
		354	136	-	5 330 162	300	increasing	110	0.3578	29.09	
		354	136	frequency	4 276 070	300	increasing	110	0.3681	26.36	
300	-	354	136	frequency/ stop words	3 091 943	300	increasing	110	0.4414	45.45	
		1 800	354	-	58 807 338	1 800	increasing	100	0.3647	18	
		1 800	354	frequency	17 588 744	1 800	80 / 0.4931	100	0.4644	40	
		1 800	354	frequency/ stop words	14 284 956	1 800	decreasing	200	0.4842	44	
		1 800	354	word classes	27 016 943	1 800	increasing	150	0.3525	68.67	
		1 800	354	word classes/frequency	11 637 733	1 800	decreasing	150	0.4984	72.67	
hermà corpus	-	1 800	354	word classes/ frequency/ stop words	11 502 025	1 800	decreasing	150	0.5018	70.67	
		1 800	354	-	51 641 735	172 810	increasing - flat	150	0.3184	36.67	
		1 800	354	frequency	35 959 559	172 810	increasing - flat	150	0.3877	58	
		1 800	354	frequency/ stop words	29 213 325	172 810	increasing - flat	100	0.4265	83	
		1 800	354	-	5 330 162	300	increasing	110	0.3578	29.09	
		1 800	354	frequency	4 276 070	300	increasing	110	0.3681	26.36	

References

- [1] D. M. Blei. “Probabilistic topic models”. In: *Communications of the ACM* 55.4 (2012), pp. 77–84. ISSN: 0001-0782, 1557-7317. DOI: 10.1145/2133806.2133826. URL: <https://dl.acm.org/doi/10.1145/2133806.2133826> (visited on 07/14/2020).
- [2] D. M. Blei, A. Y. Ng, and M. I. Jordan. “Latent dirichlet allocation”. In: *J. Mach. Learn. Res.* 3 (2003), pp. 993–1022. ISSN: 1532-4435. DOI: <http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993>. URL: <http://portal.acm.org/citation.cfm?id=944937>.
- [3] C. Blevins. *Topic Modeling Martha Ballard’s Diary*. Library Catalog: www.cameronblevins.org. 2010. URL: <https://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/> (visited on 07/14/2020).
- [4] J. Chang et al. “Reading Tea Leaves: How Humans Interpret Topic Models”. In: *NIPS*. Ed. by Y. Bengio et al. Curran Associates, Inc., 2009, pp. 288–296. ISBN: 9781615679119. URL: http://books.nips.cc/papers/files/nips22/NIPS2009_0125.pdf.
- [5] *Deutsches Textarchiv. Grundlage für ein Referenzkorporus der neuhochdeutschen Sprache*. 2020. URL: <http://www.deutschestextarchiv.de/> (visited on 07/14/2020).
- [6] G. Diaz. *German stopwords*. 2020. URL: <https://github.com/stopwords-iso/stopwords-de> (visited on 07/11/2020).
- [7] E. Gius, K. Krüger, and C. Sökefeld. “Korpuserstellung als literaturwissenschaftliche Aufgabe”. In: *DHd 2019 Digital Humanities: multimedial & multimodal Konferenzabstracts*. Frankfurt & Mainz, 2019, pp. 164–166.
- [8] B. Hermann and G. Lauer. “Das ”Was-bisher-geschah“ von KOLIMO. Ein Update zum Korpus der literarischen Moderne”. In: *DHd 2017 Digitale Nachhaltigkeit Konferenzabstracts*. Feb. 2017, pp. 107–111.
- [9] M. Jockers. “Secret” Recipe for Topic Modeling Themes. Apr. 2013. URL: <http://www.mattthewjockers.net/2013/04/12/secret-recipe-for-topic-modeling-themes/>.
- [10] M. Jockers. *The LDA Buffet is Now Open; or, Latent Dirichlet Allocation for English Majors*. Sept. 2011. URL: <https://nxnt.link/vzeXZ>.
- [11] M. L. Jockers and D. Mimno. “Significant themes in 19th-century literature”. In: *Poetics* 41.6 (2013), pp. 750–769. ISSN: 0304422X. DOI: 10.1016/j.poetic.2013.08.005. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0304422X13000673> (visited on 01/08/2020).
- [12] J. M. Lotman. *Die Struktur literarischer Texte*. München: Wilhelm Fink, 1972.
- [13] A. K. McCallum. *MALLET: A Machine Learning for Language Toolkit*. 2002. URL: <http://mallet.cs.umass.edu/>.
- [14] D. Mimno. *Package ’MALLET’*. 2015. URL: <https://cran.r-project.org/web/packages/mallet/mallet.pdf>.
- [15] A. Moles. *Informationstheorie und ästhetische Wahrnehmung*. Köln: DuMont Schauberg, 1971.
- [16] *Project Gutenberg*. en. Library Catalog: www.gutenberg.org/ (visited on 07/14/2020).

- [17] M. Röder, A. Both, and A. Hinneburg. “Exploring the Space of Topic Coherence Measures”. In: *Proceedings of the eight International Conference on Web Search and Data Mining, Shanghai, February 2-6. 2015*, pp. 399–408. URL: http://svn.aksw.org/papers/2015/WSDM_Topic_Evaluation/public.pdf.
- [18] L. V. Sacharny and A. S. Stern. “Keyword set as a type of text”. In: Lexical aspects in the system of professionally oriented training of foreign language speech activity. Perm: Perm Technical University, 1988, pp. 34–51.
- [19] C. Schöch. “Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama”. In: *Digital Humanities Quarterly* 011.2 (2017). ISSN: 1938-4122. URL: <http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html>.
- [20] A. Schofield, M. Magnusson, and D. Mimno. “Pulling Out the Stops: Rethinking Stopword Removal for Topic Models”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. EACL 2017. Valencia, Spain: Association for Computational Linguistics, 2017*, pp. 432–436. URL: <https://www.aclweb.org/anthology/E17-2069> (visited on 09/04/2020).
- [21] A. Schofield et al. “Understanding Text Pre-Processing for Latent Dirichlet Allocation”. In: *Proceedings of the 1st Workshop for Women and Underrepresented Minorities in Natural Language Processing. EMNLP 2017. 2017*, pp. 432–436. URL: <http://www.cs.cornell.edu/~xanda/winlp2017.pdf> (visited on 09/04/2020).
- [22] C. E. Shannon. “A mathematical theory of communication”. In: *Bell Syst. Tech. J.* 27.3 (1948), pp. 379–423. URL: <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>.
- [23] C. Sievert and K. Shirley. “LDAvis: A method for visualizing and interpreting topics”. In: *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. Baltimore, Maryland, USA: Association for Computational Linguistics, 2014*, pp. 63–70. DOI: 10.3115/v1/W14-3110. URL: <http://aclweb.org/anthology/W14-3110> (visited on 05/02/2020).
- [24] S. Syed and M. Spruit. “Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation”. In: *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA). 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA). Tokyo, Japan: IEEE, 2017*, pp. 165–174. ISBN: 978-1-5090-5004-8. DOI: 10.1109/DSAA.2017.61. URL: <http://ieeexplore.ieee.org/document/8259775/> (visited on 05/31/2020).
- [25] *TextGrid Repository*. 2020. URL: <https://textgridrep.org/browse/root> (visited on 07/14/2020).
- [26] T. Underwood. *Topic modeling made just simple enough*. Apr. 2012. URL: <https://nxnt.link/g4D6p>.
- [27] T. Underwood. *What kinds of “topics” does topic modeling actually produce?* Apr. 2012. URL: <https://nxnt.link/g4D6p>.
- [28] T. Weitin and K. Herget. *Falcon Topics*. 4. Digital Humanities Cooperation, 2016. 1-20. URL: <https://www.digitalhumanitiescooperation.de/pamphlete/pamphlet-4-falkentopics/>.