CHAPTER 15

The Many Themes of Humanism

Topic Modelling Humanism Discourse in Early 19th-Century German-Language Press

Heidi Hakkarainen and Zuhair Iftikhar

Introduction

Topic modelling is often described as a text-mining tool for conducting a study of hidden semantic structures of a text or a text corpus by extracting topics from a document or a collection of documents.[1] Yet, instead of one singular method, there are various tools for topic modelling that can be utilised for historical research. Dynamic topic models, for example, are often constructed temporally year by year, which makes it possible to track and analyse the ways in which topics change over time.[2] This chapter provides a case example on topic modelling historical primary sources. We are using two tools to carry out topic modelling, MALLET and Dynamic Topic Model (DTM), in one dataset, containing texts from the early 19th-century German-language press which have been subjected to optical character recognition (OCR). All of these texts were discussing humanism, which was a newly emerging concept before mid-century, gaining various meanings in the public discourse before, during and after the 1848–1849 revolutions. Yet, these multiple themes and early interpretations of humanism in the press have been previously under-studied. By

How to cite this book chapter:
Hakkarainen, H., & Iftikhar, Z. (2020). The many themes of humanism: Topic modelling humanism discourse in early 19th-century German-language press. In M. Fridlund, M. Oiva, & P. Paju (Eds.), *Digital histories: Emergent approaches within the new digital history* (pp. 259–277). Helsinki: Helsinki University Press. https://doi.org/10.33134/HUP-5-15

analysing the evolution of the topics between 1829 and 1850, this chapter aims to shed light on the change of the discourse surrounding humanism in the early 19th-century German-speaking Europe.

The concept of humanism (*Humanismus*) was first coined by Friedrich Immanuel Niethammer (1766–1848) in German-speaking Europe in 1808.[3] The concept was originally used in the pedagogical debate concerning education, especially in the *Gymnasium*. This pedagogical debate between humanist and philanthropist (realist) education was related to 19th-century educational reforms and especially to the school reform in Bavaria, which preceded the Prussian school reform between 1809 and 1819.[4] However, in addition to these pedagogical debates, the concept of humanism spread more widely in the 1830s and 1840s, and in this gained new meanings and interpretations.

However, as the previous studies have focused on the early 19th-century pedagogical debates, this wider dissemination and popularisation of the new concept in the printed press has not been under an extensive close study.[5] There exist a large number of printed publications discussing humanism already in the first part of the 19th century, which makes an inquiry on press debates a challenging task for historians.[6] In order to tackle this challenge of the vast size of potential source material, this chapter uses the quantitative method of computer-based 'topic modelling' to assist the qualitative analysis.

By topic modelling a set containing almost 100 key texts from the years between 1829 and 1850, this chapter recovers several of those multiple discourses connected to humanism before, during and after the outburst of the 1848–1849 revolutions. By combining and comparing topic modelling with MALLET with Dynamic Topic Modelling (DTM), this chapter seeks to map and analyse what kinds of topics were related to humanism before 1850 and how these topics changed and evolved over time. During 21 years, humanism appeared in various contexts from education to philosophy, religion and politics. Where the MALLET, as the most well-established topic modelling tool within the field of digital history,[7] is used in detecting the most prominent themes in the discussion on humanism, DTM makes available a finer look into the topics at the temporal level and, in this case study, provides a new kind of insight into the growing importance of temporality within the German-language humanism discourse between the early 1800s and the mid-century.[8]

In contrast to temporally ambitious research on huge corpuses, this chapter focuses on a rather small text corpus, which allows more exploration of the possibilities of cross-reading the material with methods of close and distant reading. This study of the discussion surrounding humanism before 1850 thus provides a reasonably manageable but rich investigation of some of the ways in which newspapers and periodicals addressed topical issues and transferred concepts and new ideas across political borders within the lands of the German Confederation.

At the same time, we seek to explore what kinds of methodological benefits and risks are involved in the topic modelling of historical sources. The technique of topic modelling decides what constitutes a topic on an algorithm that creates a statistical model of word clusters. It is thus not a fixed schema, but a variable probabilistic model that should also be treated as such. We will demonstrate how various forms of cleaning and filtering of the data can have drastic results on the output of the topic model. We also present and compare outputs from different methods of topic modelling, using the MALLET application and DTM, and address various methodological concerns related to topic modelling.

Topic Modelling with MALLET

The first essential step in describing the 19th-century German-language press discourse on humanism was to identify its various individual themes or topics using the quantitative method of topic modelling. Topic modelling has its roots in information retrieval, natural language processing and machine learning. This probabilistic tool has attracted attention among historians, because it enables detecting underlying thematic structures behind a large corpus of documents, as well as surprising connections between individual texts. A topic comprises a distribution of words. A single document is assumed to contain words about multiple topics within the whole dataset. Each word is drawn from those dataset topics. The study used two topic modelling tools where the first is called MALLET (Machine Learning for Language Toolkit, version 2.0.8.), which is an open source Java-based software package for natural language processing using Latent Dirichlet Allocation technique (LDA).[9]

Before using MALLET and in pre-preparation, the machined encoded OCR German-language press texts were cleaned and corrected manually (especially the recurring problem with some Unicode characters). In some cases, this included shortening the texts by excluding clearly irrelevant sections.

The model was then made with the 'optimise-interval' command, which sets each topic's probabilistic Dirichlet parameter that indicates the topic's proportion in the whole dataset, and gives a better fit to the data by allowing some topics to be more prominent. In addition, the number of topics to be identified by MALLET is set beforehand as there is no 'natural' number of topics in a corpus, but this part requires manual evaluation and iteration by the researchers.[10] Both MALLET and the DTM tool only mechanically detect topics and assign them numeric values, whereas identifying and naming the topics (that is, determining and labelling the thematic categories found by the machine reading) is something the human researcher has to carry out using manual reading. And this is an act of interpretation.

Topic Modelling with DTM

Within probabilistic topic modelling, LDA is a frequently used technique and its MALLET implementation has traditionally been the most popular tool to analyse historical corpuses. Ever since topic modelling was first introduced in the early 2000s, there have been new extensions that help to model temporal relationships. One shortcoming of the LDA method is that it assumes that the order of documents is irrelevant. But if we – as historians are often prone to – want to discover the evolution of topics over time, then we have to take the time sequence into account. DTM attempts to overcome this shortcoming and captures the dynamics of how topics emerge and change over time.[11]

DTM is designed to explicitly model the ways in which topics evolve over time and to give qualitative insights into the changing composition of the source material. However, it is not the only such tool available and it has also been subjected to critique for penalising large changes from year to year.[12] The DTM is a probabilistic time series model, which is designed to track and analyse the ways in which latent topics change over time within a large set of documents. For example, David M. Blei and John D. Lafferty demonstrated the functioning of DTM by investigating topics of the journal *Science* between 1880 and 2000.[13] Our case study is based on a small source corpus, which, as we will soon see, was one important factor in the output from the dynamic topic modelling. Because of the small size of the dataset, cleaning and filtering the data had a major impact on DTM's output. The more historical sources were pre-processed, the more stable the model became.

As mentioned above, few but not all text files were reviewed for common mistakes and in a few instances some mistakes were manually corrected. Python's Natural Language Toolkit (nltk) library was used for the pre-processing and filtering of the texts. Prior to passing on the text data to DTM tools, the text was processed using the following pre-processing pipeline:

1. **Punctuation and numbers removal.** Punctuation characters within and around all the words were deleted and all the other characters except alphabetic characters were removed.
2. **Stop words removal.** This is a common operation when processing text in any domain. The list of German stop words was initially taken from the nltk library and MALLET tool. This list was extended by reviewing the texts and some words deemed to be useless were then added to the list. Any words in the stop words list were removed in pre-processing.
3. **Stemming and lemmatising.** Stemming is the process by which a word is reduced to its base form and all the inflectional forms of a word are reduced to a single base stem. Using language dictionaries, lemmatisation converts a word to its base lemma. This is the word from which all the inflectional forms are derived. The base stem is then used by the lemmatisers to find the base lemma, which is then kept in the text.
4. **Classification.** The words were then classified into different parts-of-speech with the goal being to keep various nouns and verbs identified in the input texts. Words which belong to other parts-of-speech were removed.
5. **Rare words.** As a final step, the words which appear only once in the whole input corpus were also removed.

We then used Gensim (Python library) to run the DTM tool. After creating various outputs of models with 5 to 20 topics, as for the previous analysis using MALLET, we decided to limit the number of topics to 10. Like MALLET, DTM also gives keywords (that is, a cluster of words relevant to the topic), which help to identify the topic.

Source Material

The source material used in this study is a sub-dataset from the digital corpus Austrian Newspapers Online (ANNO), provided by the Austrian National Library (at http://anno.onb.ac.at). The digital ANNO collection contains around 20 million pages of German-speaking newspapers and periodicals that are available for full text searches.[14] The Austrian National Library at their ANNO-portal provides an OCR tool for machine encoded optically recognised text which, although not totally reliable and contains errors, can be used for the digital analysis.

According to the full text search engine of the ANNO portal, the word *Humanismus* (humanism) was mentioned 326 times in the press between 1808 and 1850.[15] Because the old German *Fraktur* typeface is challenging for OCR, the results should not be interpreted as entirely reliable, but as giving an indication of the scale of use, how much this word was circulated in the press. In some texts, humanism appeared only once in passing, while in others it was mentioned several times and discussed explicitly. Based on their relevance, length and readability, we have selected 95 key texts for topic modelling analysis (see Appendix 15.1). These texts include book reviews, articles, news, feuilleton writings and political reports, while reprints, short notices, adverts and obituaries have been excluded.

Figure 15.1 illustrates the publishing centres and various publications that make up the dataset. The graph is made with the Gephi visualisation application and it aims to depict the source material in a visually conceivable way. Moreover, Gephi is a frequently used software tool for network analysis, because it enables the portrayal and analysis of relationships or interaction between persons, entities and objects, such as geographical places or publications.[16] The objects (nodes) and their relationships (edges) can be presented in many different ways. In this case, the layout was made manually instead of choosing one of the most popular layout algorithms such as Force Atlas or Fruchterman Reingold. The nodes and edges tables were imported to Gephi as CSV files and in the edges table the connection between a publication and its place of publishing gained 'weight' in accordance to the amount of texts discussing humanism in that particular publication during the period between 1829 and 1850. The more humanism was mentioned, the thicker the line between a newspaper or a magazine and the city in which it was published. Accordingly, the strength of connections indicates which were the most important publishing centres

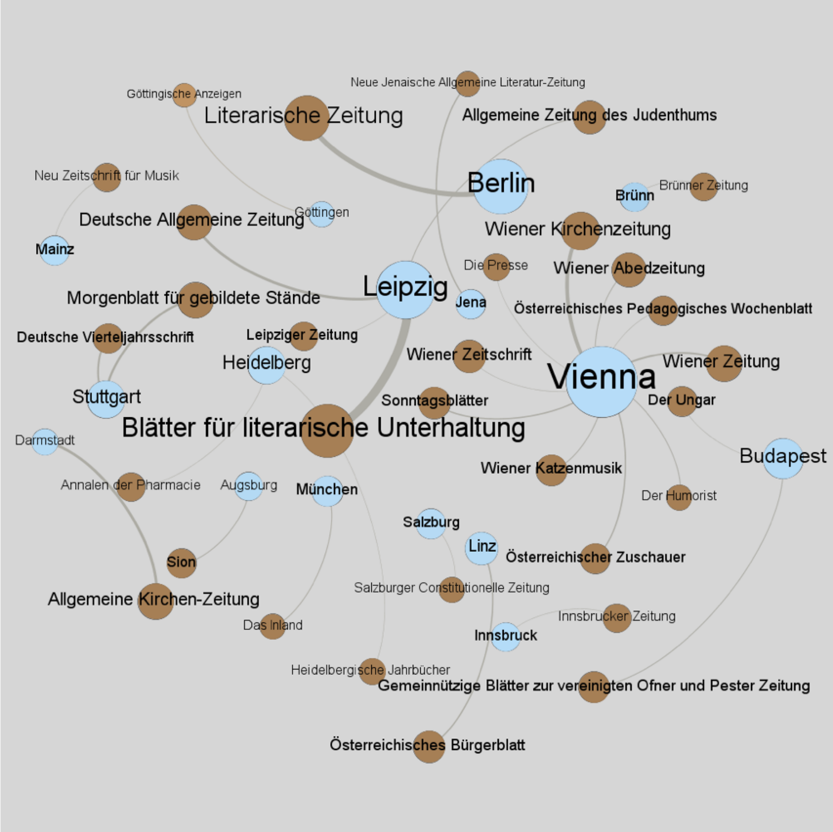

Neue Jenaische Allgemeine Literatur-Zeitung
Göttingische Anzeigen
Literarische Zeitung
Allgemeine Zeitung des Judenthums
Neu Zeitschrift für Musik
Berlin
Brünner Zeitung
Brünn
Deutsche Allgemeine Zeitung
Göttingen
Wiener Kirchenzeitung
Mainz
Die Presse
Wiener Abedzeitung
Morgenblatt für gebildete Stände
Leipzig
Jena
Österreichisches Pedagogisches Wochenblatt
Deutsche Vierteljahrsschrift
Leipziger Zeitung
Heidelberg
Wiener Zeitschrift
Vienna
Wiener Zeitung
Stuttgart
Sonntagsblätter
Der Ungar
Darmstadt
Blätter für literarische Unterhaltung
Budapest
Wiener Katzenmusik
Annalen der Pharmacie
Augsburg
München
Der Humorist
Salzburg
Linz
Österreichischer Zuschauer
Sion
Allgemeine Kirchen-Zeitung
Salzburger Constitutionelle Zeitung
Innsbrucker Zeitung
Das Inland
Innsbruck
Heidelbergische Jahrbücher
Gemeinnützige Blätter zur vereinigten Ofner und Pester Zeitung
Österreichisches Bürgerblatt

Figure 15.1: Network diagram of publications and publishing centres of the dataset. Source: Authors.

and highlights publications that most extensively dealt with humanism in this dataset. Even though the ANNO source corpus is partial and dominated by Austrian newspapers and magazines, Figure 15.1 shows that the early 19th-century discussion on humanism surpassed political borders within the German Confederation spreading in the area of fragmented German lands and German-speaking parts of the Habsburg Empire. Vienna, Leipzig and Berlin were the most important publishing centres and the literary journals *Blätter für literarische Unterhaltung* and *Literarische Zeitung* dealt with the topic most extensively, although the publications dealing with humanism ranged from daily newspapers to religious magazines and satirical journals.

German Humanism According to MALLET Topic Modelling

Initial details about MALLET are summarised in the previous section. Below are the eight topics in order of prevalence with their top words as discovered by MALLET when asked to determine the 10 most prevalent topics and as labelled (education, reformation, etc.) by us. The number of topics was chosen after experimenting with different kinds of models and 10 topics were chosen as a best way for modelling the source corpus, which was small and fragmented. Topic modelling usually involves filtering away so-called stop words, non-informative frequently appearing words such as articles, particles and pronouns. However, especially when it comes to creating a model with a small number of topics, pre-processing the data has a danger of compromising the results as the researcher makes decisions on removal of stop words according to her or his pre-understanding, thus projecting into the data certain presuppositions regarding what is important in the corpus.[17] Accordingly, in this model, no pre-filtering of stop words was carried out before the analysis, but two topics that contained only stop words were filtered out after creating the model. See Appendix 15.2 for the whole model.

Religion: fich menschen gott religion find juden zukunft religiösen gottes humanismus mensch christenthum christliche niht darum demokratie humanität christlichen christen theorie

Education: erziehung schulen lehrer sprache bildung seyn gymnasien unterricht realismus sprachen realschulen schüler jugend individuum wissenschaften anstalten schrift realschule

Revolution: wurde freiheit volk stadt wurden berlin revolution kammer bald volkes völker waren heute republik straßen preußen fast macht bürgerwehr haufen

Philosophy: fich philosophie ruge find nationalismus princip paris jahrbücher literatur preußen geschrieben briefe socialismus anfichten brief patriotismus rage artikel principien staatsanwalt

Reformation: kirche fich universitäten luther reformation staat lehre staats reform gemeinden schottischen glaubens bloß kirchen verfassung staate theologen lehrer wissenschaft hervor

Death penalty: todesstrafe sei abg verbrechen strafe habe amendement antrag könne man dieß gesetze redner verbrecher abgeschafft jury wolle abschaffung angenommen gegen

Press debate: daſs christlichen philologie gegner muss zeitung liberalismus sache sinne bedeutung gesinnung jedenfalls artikel presse giebt philologen meinung klassischenmonarchischen christliche

Social issues: ſie the ſich hamburg euch gesehen zigeuner habt bey iſt dieſe wiſſen ſeine sprachen stadt armen glück schüler jhr their

The output from MALLET provides eight topics with different keywords. In the ‘Education’ topic, words like *Erziehung* (education/upbringing), *Schulen* (schools), *Lehrer* (teacher) and *Sprache* (language) are clustered together with such difficult-to-translate German concepts like *Bildung* and *Gymnasien*, which indicate that this topic is related to the educational debates about the role of humanism in the modern schooling system that were a very important issue in the era of comprehensive school reforms. After all, the concept of

Humanismus (humanism) was, as mentioned above, first coined as a pedagogical concept, fostering classical education and the study of classical languages.[18] The 'Reformation' topic, on the other hand, contains words like *Kirche* (church), *Universitäten* (universities), *Luther* (Luther) and *Reformation* (reformation), which give reason to believe that this topic deals with humanism historically in relation to Martin Luther and the reformation era.

However, in addition to these highly obvious and clear results, there are also topical word clusters which show a completely different kind of interpretation of humanism. The topic 'Philosophy', for instance, contains words like *Philosophie* (philosophy), *Ruge* (Ruge), *Nationalismus* (nationalism), *Princip* (principle) and *Paris* (Paris). All of these words are connected to the philosopher Arnold Ruge (1802–1880), who was also a political writer, associated with the Young Hegelians and Karl Marx, and known for his radical ideas that religion should be separated from politics and intellectual thinking. Ruge was one of the main figures who in the 1840s introduced a new interpretation of humanism as a political concept and his ideas were highly debated in the press.[19] For Ruge, humanism meant political emancipation from the old *ancien régime*. He incorporated humanism in democratic-republican ideology, which combined social critique with critique towards religion and growing nationalism. Humanism meant political, religious and social freedom, which was universal for the whole of mankind and superseded national borders. Accordingly, in *Geschichtliche Grundbegriffe*, Ruge's interpretation of humanism is called *kosmopolitischer Humanismus* (cosmopolitan humanism).[20]

This radical new political meaning of the concept of humanism is also visible in topics that dealt with social problems and political issues like the death sentence and the 1848–1849 revolution. For example, the topic labelled 'Social issues' contains keywords like *Zigouner* (gypsies), *Armen* (the poor), *Stadt* (city) and *Glück* (happiness). Again, the topic 'Death penalty' is clustering together words like *Todesstrafe* (death penalty), *Verbrechen* (felony), *Strafe* (punishment) and *Amendement* (amendment), which are all related to the debates around abolishment of the death penalty, which was a topical issue especially in Austria around 1849. Moreover, topic modelling of the dataset reveals a topic relating explicitly to the European revolutions in 1848–1849. This topic labelled with the title 'Revolution' contains the following keywords: *wurde* (came), *Freiheit* (freedom), *Volk* (people), *Stadt* (city), *Berlin* (Berlin), *Revolution* (revolution), *bald* (soon), *heute* (today), *Republik* (republic) *Straßen* (streets), *Macht* (power), *Bürgerwehr* (militia) and *Haufen* (pile). This topic, especially, indicates how humanism became a political concept in the 1840s when both early socialists and liberals adopted humanism in their political language as they demanded political emancipation from the old regime.[21]

This result demonstrates the diversity of the meanings given to humanism in the early 19th-century press. In addition to educational debates, humanism also appeared in the discussions surrounding social and moral issues, law and politics. In fact, the extremely diverse topics of humanism indicate a pervasive reorganising of ideas related to the human being and his or her place in the universe in the post-Napoleonic era, in which the liberal bourgeoisie was gaining a new foothold in society at the same time that the Church and absolutist power were challenged in the aftermath of the French Revolution. This transformative era created new interpretations on how politics, religion, education and philosophical thinking should be organised in modern secularising society, and, despite the practices of censorship especially in Prussia and Austria,[22] the press played a major role in circulating these ideas among a growing readership.

Consequently, the vast processes of secularisation and modernisation help us to understand why the 'Religion' topic was the most dominant theme in the early 19th-century press discussion on humanism. This most prevalent topic contains many interesting keywords indicating how discourses surrounding religion, morality and politics were actually significantly entangled in the early 19th-century discussion on humanism. The clustering of words like *Menschen* (human being), *Gott* (god), *Religion* (religion), *Juden* (Jews), *Zukunft* (future), *Humanismus* (humanism), *Christenthum* (Christianity), *Demokratie* (democracy), *Humanität* (humanity) and *Theorie* (theory) is a good example of the interpretative challenges that take place when identifying and labelling topics that are not cohesive but multifaceted and extremely complex. We will examine the 'Religion' topic closer below using DTM. But first, we will locate which years this topic emerged most dominantly between 1829 and 1850.

Following the task of identifying topics, it is vital to also explore them and their meanings in the historical context in which they came to life. In other words, it is essential to acknowledge the temporality of the topics and study them from in a dynamic historical perspective. For example, the volume of the press was very different in 1829 and in 1850. Furthermore, the new Young Hegelian philosophical ideas and growing interest in social issues was part of the intellectual and social landscape of the 1840s and it goes without saying that the outbreak of the revolutions in 1848 was clearly a major historical event that impacted on the public discourse surrounding humanism.

Without additional programming, MALLET does not present the topics in relation to time. Yet, it is possible to inspect the dynamic temporal aspect of the topics by organising the dataset chronologically.[23] Accordingly, the files of the dataset were numbered from the oldest, in this case 1829, to the youngest, here 1850. This means that it is now possible to study how topics emerged and changed over time (Figure 15.2). In Figure 15.2, the two stop word topics are filtered out, presenting only the eight relevant topics.

We can now see the thematic trends and how the topic patterns change over time. The figure above indicates that before 1840 'Education' and 'Social issues' were important topics in relation to humanism, but in 1848 the topic 'Revolution' became dominant. In 1849, it was replaced as the leading topic by 'Death penalty', with 'Religion' following in prevalence. The 'Religion' topic gained importance especially immediately after the revolution, which could indicate a reaction to the turbulence and violence in 1848–1849. Yet, despite

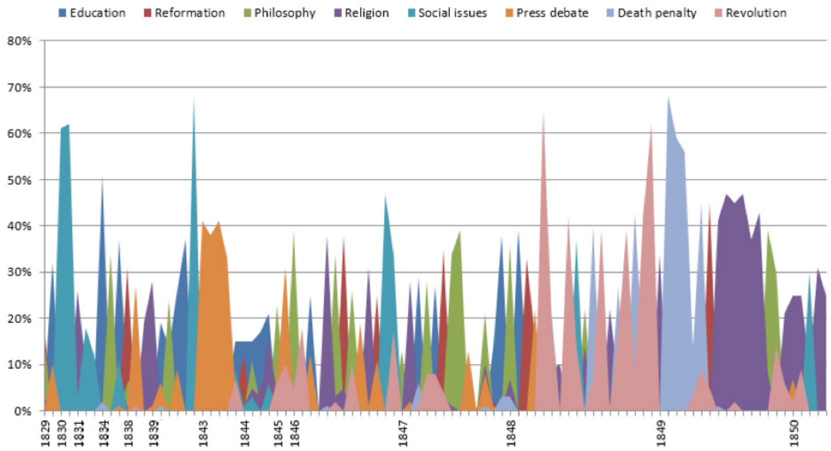


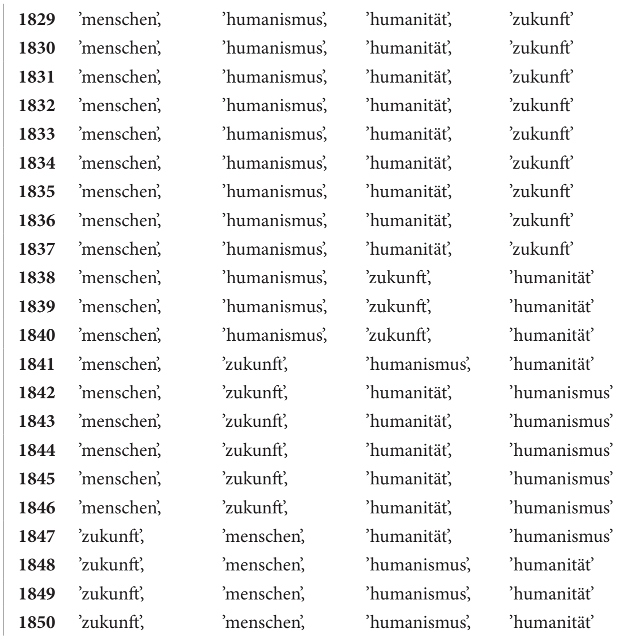
Figure 15.2: Annual allocation of topics. Source: Authors.

the chronological aspect, MALLET's results are always compressed and cannot give any further insight into the dynamics within the topics that have been discovered. In the next section, we will further analyse how dynamic topic modeling (DTM) can make it possible to gain insight about the dynamics within one singular topic.

Discovering Temporalisation of German Humanism with DTM

Preliminary details about DTM and text pre-processing details are mentioned above. The cleaning of the data had a major impact on the output of the model. At first, the results were very similar to the MALLET analysis and many topics seen before persisted. For example, humanism continued to emerge in relation to the topic of 'Religion' ['menschen', 'humanismus', 'humanität', 'zukunft', 'ste', 'stch', 'religion', 'wahrheit', 'bloß', 'demokratie', 'christenthum', 'recht', 'gegenwart', 'fich', 'wohl']. Also, the topics 'Education' ['bildung', 'mehr', 'erziehung', 'zeit', 'lehrer', 'jugend', 'realschulen', 'find', 'wissenschaft', 'sache', 'neuen', 'immer', 'gymnasien', 'zweck', 'mittel'], 'Death penalty' ['sei', 'verbrechen', 'abg', 'könne', 'redner', 'schon', 'antrag', 'amendement', 'angenommen', 'staat', 'dieß', 'abgeschafft', 'ab', 'be', 'abschaffung'] and 'Revolution' ['wurde', 'macht', 'völker', 'volk', 'geschichte', 'bald', 'freiheit', 'berlin', 'volkes', 'revolution', 'werk', 'wurden', 'je', 'regierung', 'tage'] were remarkably similar. However, there were also changes. Social issues and debates around Ruge's interpretation of humanism were more in the background and there was more than one category relating to religion and education.

Furthermore, with DTM, we had more fine-tuned results as the source corpus was divided into different time frames and keywords were arranged year by year. As the keywords appeared in a list from most important to least



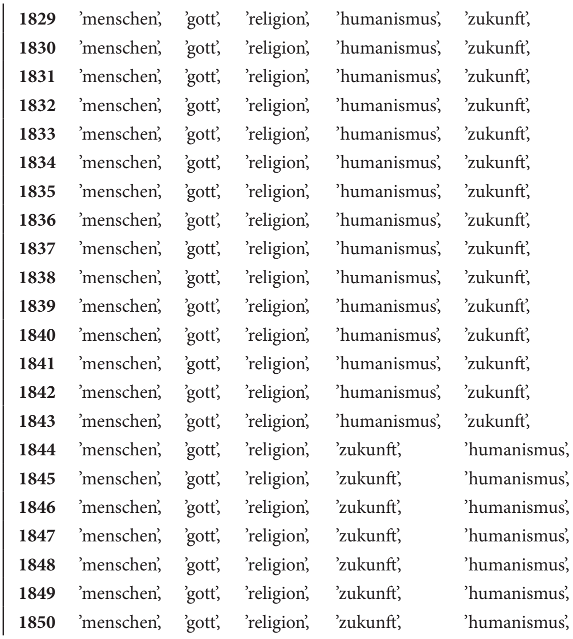
```
1829  ’menschen’,   ’humanismus’,  ’humanität’,   ’zukunft’
1830  ’menschen’,   ’humanismus’,  ’humanität’,   ’zukunft’
1831  ’menschen’,   ’humanismus’,  ’humanität’,   ’zukunft’
1832  ’menschen’,   ’humanismus’,  ’humanität’,   ’zukunft’
1833  ’menschen’,   ’humanismus’,  ’humanität’,   ’zukunft’
1834  ’menschen’,   ’humanismus’,  ’humanität’,   ’zukunft’
1835  ’menschen’,   ’humanismus’,  ’humanität’,   ’zukunft’
1836  ’menschen’,   ’humanismus’,  ’humanität’,   ’zukunft’
1837  ’menschen’,   ’humanismus’,  ’humanität’,   ’zukunft’
1838  ’menschen’,   ’humanismus’,  ’zukunft’,     ’humanität’
1839  ’menschen’,   ’humanismus’,  ’zukunft’,     ’humanität’
1840  ’menschen’,   ’humanismus’,  ’zukunft’,     ’humanität’
1841  ’menschen’,   ’zukunft’,     ’humanismus’,  ’humanität’
1842  ’menschen’,   ’zukunft’,     ’humanität’,   ’humanismus’
1843  ’menschen’,   ’zukunft’,     ’humanität’,   ’humanismus’
1844  ’menschen’,   ’zukunft’,     ’humanität’,   ’humanismus’
1845  ’menschen’,   ’zukunft’,     ’humanität’,   ’humanismus’
1846  ’menschen’,   ’zukunft’,     ’humanität’,   ’humanismus’
1847  ’zukunft’,    ’menschen’,    ’humanität’,   ’humanismus’
1848  ’zukunft’,    ’menschen’,    ’humanismus’,  ’humanität’
1849  ’zukunft’,    ’menschen’,    ’humanismus’,  ’humanität’
1850  ’zukunft’,    ’menschen’,    ’humanismus’,  ’humanität’
```

Figure 15.3: Output from the DTM before data cleaning, including the four first keywords. Source: Authors.

important, it was possible to detect the ways in which the order of these keywords changed within one singular topic. The most striking new discovery with DTM was that there were cases in which words with temporal meaning such as *Zeit* (time) or *Zukunft* (future) became increasingly important towards mid-century. This discovery resonates strongly with the conceptual historian Reinhart Koselleck’s famous argument that the early 19th century was a *Sattelzeit*, a period in which the notion of time changed radically and concepts became increasingly abstract and more future-oriented. Koselleck suggested that as modern concepts became more entangled with historical time, being associated increasingly with the past, the present and the future, the phenomena which previously were seen as static and unchanging became conceived as dynamic processes.[24]

To give an example, in Figure 15.3 we have the four most important words for the topic ‘Religion’, containing words like *Menschen* (human being), *Humanismus* (humanism), *Zukunft* (future), *Humanität* (humanity), *Religion* (religion), *Wahrheit* (truth), *Demokratie* (democracy), *Christenthum* (Christianity), *Recht* (justice) and *Gegenwart* (present), which are very similar to those words seen in the most prevalent ‘Religion’ topic in the MALLET results.

However, here the topic seems to be relating more to human beings and morality rather than religion. In addition, the meaning of the word *Zukunft* (future) is of special interest here, as its position changes radically between



```
1829  ’menschen’,  ’gott’,  ’religion’,  ’humanismus’,  ’zukunft’,
1830  ’menschen’,  ’gott’,  ’religion’,  ’humanismus’,  ’zukunft’,
1831  ’menschen’,  ’gott’,  ’religion’,  ’humanismus’,  ’zukunft’,
1832  ’menschen’,  ’gott’,  ’religion’,  ’humanismus’,  ’zukunft’,
1833  ’menschen’,  ’gott’,  ’religion’,  ’humanismus’,  ’zukunft’,
1834  ’menschen’,  ’gott’,  ’religion’,  ’humanismus’,  ’zukunft’,
1835  ’menschen’,  ’gott’,  ’religion’,  ’humanismus’,  ’zukunft’,
1836  ’menschen’,  ’gott’,  ’religion’,  ’humanismus’,  ’zukunft’,
1837  ’menschen’,  ’gott’,  ’religion’,  ’humanismus’,  ’zukunft’,
1838  ’menschen’,  ’gott’,  ’religion’,  ’humanismus’,  ’zukunft’,
1839  ’menschen’,  ’gott’,  ’religion’,  ’humanismus’,  ’zukunft’,
1840  ’menschen’,  ’gott’,  ’religion’,  ’humanismus’,  ’zukunft’,
1841  ’menschen’,  ’gott’,  ’religion’,  ’humanismus’,  ’zukunft’,
1842  ’menschen’,  ’gott’,  ’religion’,  ’humanismus’,  ’zukunft’,
1843  ’menschen’,  ’gott’,  ’religion’,  ’humanismus’,  ’zukunft’,
1844  ’menschen’,  ’gott’,  ’religion’,  ’zukunft’,     ’humanismus’,
1845  ’menschen’,  ’gott’,  ’religion’,  ’zukunft’,     ’humanismus’,
1846  ’menschen’,  ’gott’,  ’religion’,  ’zukunft’,     ’humanismus’,
1847  ’menschen’,  ’gott’,  ’religion’,  ’zukunft’,     ’humanismus’,
1848  ’menschen’,  ’gott’,  ’religion’,  ’zukunft’,     ’humanismus’,
1849  ’menschen’,  ’gott’,  ’religion’,  ’zukunft’,     ’humanismus’,
1850  ’menschen’,  ’gott’,  ’religion’,  ’zukunft’,     ’humanismus’,
```

Figure 15.4: Output from the DTM after data cleaning, including the five first keywords. Source: Authors.

1829 and 1850. Figure 15.3 shows the output from the DTM before data cleaning, including the four first keywords. In post-cleaning, the letters ‘ste’ were filtered out.

However, this striking change did not appear in all the outputs, but the more we removed stop words and filtered the data for better results, the more stable the topic appeared (Figure 15.4). In addition, the word God (*Gott*), which is missing in the first output together with religion (*Religion*), is now continuously the second most important word after human being (*Mensch*). The information about the proposition of each word within the topic indicates that changes were so minor that altering the script by removing stop words and removing words that appeared only once changed and stabilised the model to the extent that changes could no longer be seen in the order of the keywords.[25]

Yet, to give another example, the word *Zeit* (time) became increasingly important in another topic that included keywords such as *Wissenschaft* (science/knowledge) and *Erziehung* (education/upbringing). The change is visible both before and after filtering stop words. The Dirichlet parameter indicates that the weight of the word *Zeit* did not increase, but the growing importance resulted from the fact that the importance of the word *Wissenschaft* decreased radically around 1846.[26] This was a modest change, but it persisted in the outputs made before and after removing the stop words and carrying out other data filtering, such as removing words that appeared only once.

1829–1850
Philology
Church history
Philosophical tendencies
Revolution & Distribution of power
Political debate
Philosophy of science
Education of the Jews
Education
Religion, Morality & Relationship to God
Study of languages

Figure 15.5: Topics detected by the DTM tool after data cleaning. Source: Authors.

In the end, after data cleaning and filtering the historical sources, the DTM tool provided a list of the 10 most prevalent topics in the early 19th-century press (Figure 15.5). Yet, because of the short timeline and small size of the source corpus, the final output provided very static results and only very small changes within these topics were able to be discovered.

However, it is important to bear in mind that the dataset used in this case study was small. A larger dataset together with a potentially longer timeline would probably make it possible to detect and analyse more drastic changes over time. In any case, both of these examples illustrate that topic models are first and foremost probabilistic models providing estimates of the most salient discourse topics. Semantic changes are related to probabilistic proportional changes (in topic word list) and examining the probability distribution parameters (values associated with topic words in the output) is vital for understanding how these models work in practice.

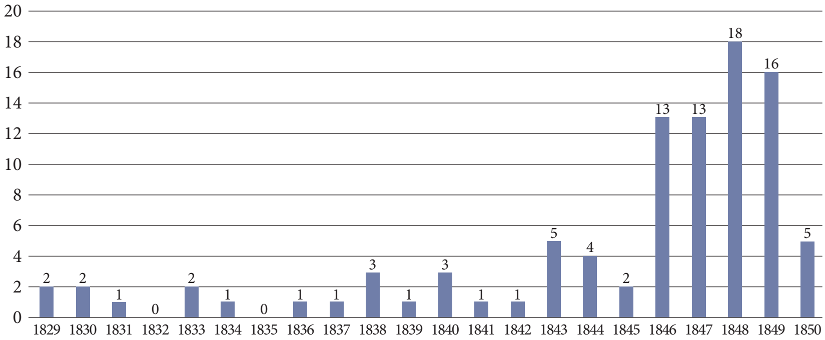
Conclusions

This study has investigated the early 19th-century German press discourse on humanism, which has been an under-researched area to date. In this chapter, we have modelled the topics of humanism in the early 19th-century German-language press with MALLET and DTM. By analysing the evolution of the topics between 1829 and 1850, this chapter has explored the change of the discourse surrounding humanism in early 19th-century German-speaking Europe. Both topic modelling applications detected different topics among the text corpus and recognised different semantic categories in the early 19th-century German-language source material without any understanding of the substance or context of these texts.

Topic modelling contains various methods, which can be used for different purposes. As we have shown, topic modelling can provide assistance for historical research as a tool for analysis and interpretation. In this study, we created different topic models of a dataset that was relatively small and could be closely read in addition to distant reading. Both MALLET and the DTM tool not only enable us to identify thematic categories (that is, topics within the dataset), but they also make it possible to trace these topics back to file level. The outputs produced detailed results on how each topic appeared in each of the 95 texts of the dataset, which makes it possible to trace topics back to the level of individual articles for close reading analysis. If one is especially interested in, say, 'Revolution' as a press topic, one could select and read all the news articles and other texts in which this topic appeared during the time frame of 1829 to 1850. This kind of assistance is invaluable for mapping and assessing sources, which is often laborious and time-consuming.

At the same time, our study also sheds light on the potential benefits and risks of topic modelling within historical research. From a methodological perspective, it is important to bear in mind that although topic modelling might produce highly compelling results, the analysis of these results demands time, skills and caution. One has to remember that results can vary depending upon the input topic number, size of the dataset, specific tool used for topic modelling, data cleaning and methods of filtering. Topic modelling provides assistance for historical research as a tool for analysis and interpretation, but the output of a topic modelling process is not a result in itself and needs to be studied further for reliable conclusions. Topic modelling results can answer a historian's intuitive questions by providing focus and direction to the analysis of historical corpuses through traditional methods of historical inquiry, source criticism, close reading and contextualisation. Perhaps even more importantly, topic modelling has the potential to challenge established patterns of thought and underlying presumptions by providing a completely different angle on historical sources.

Appendix 15.1: Dataset



Appendix 15.2: MALLET Tool

Topic ID	Topic label	Dirichlet parameter	Keywords
0	[EDUCATION]	0,12609	erziehung schulen lehrer sprache bildung seyn gymnasien unterricht realismus sprachen realschulen schüler jugend individuum wissenschaften anstalten schrift realschule
1	[REFORMATION]	0,07934	kirche fich universitäten luther reformation staat lehre staats reform gemeinden schottischen glaubens bloß kirchen verfassung staate theologen lehrer wissenschaft hervor
2	[PHILOSOPHY]	0,09784	fich philosophie ruge find nationalismus princip paris jahrbücher literatur preußen geschrieben briefe socialismus anfichten brief patriotismus rage artikel principien staatsanwalt
3	[RELIGION]	0,18337	fich menschen gott religion find juden zukunft religiösen gottes humanismus mensch christenthum christliche niht darum demokratie humanität christlichen christen theorie
4	[SOCIAL ISSUES]	0,05893	ſie the ſich hamburg euch gesehen zigeuner habt bey iſt dieſe wiſſen ſeine sprachen stadt armen glück schüler jhr their
5	[STOP WORDS]	1,43466	mit aber nur man hat noch diese zeit welche haben mehr gegen denn selbst uns alle ohne ihm sondern leben
6	[PRESS DEBATE / CONTROVERSY	0,06722	daſs christlichen philologie gegner muss zeitung liberalismus sache sinne bedeutung gesinnung jedenfalls artikel presse giebt philologen meinung klassischen monarchischen christliche
7	[LAW / DEATH PENALTY]	0,07172	todesstrafe sei abg verbrechen strafe habe amendement antrag könne man dieß gesetze redner verbrecher abgeschafft jury wolle abschaffung angenommen gegen
8	[REVOLUTION]	0,11352	wurde freiheit volk stadt wurden berlin revolution kammer bald volkes völker waren heute republik straßen preußen fast macht bürgerwehr haufen
9	[STOP WORDS]	0,07725	fich ift find feine fein diefe diefer feiner fei fondern nnd felbfi ihm nichts zwifchen diefem fchon lehre fehr wol

Appendix 15.3: DTM Tool

10 topics found by DTM after data cleaning:

Philology: philologie alterthums wissenschaft studien zeit bedeutung artikel klassischen richtung gymnasien weise zeigt damals gewinnen darauf

Church history: kirche deutschen humanismus ganz zeit Deutschland große ruge macht geschichte staat bald reformation freiheit princip hätte jahrhunderts

Philosophical tendencies: tendenz so deutfchen bey welt bildet herr wißen vermögen gerade briefe feuerbach menfchen diese

Revolution & Distribution of power: schon geht volk bleibt freiheit berlin verbrechen hand viele ersten fall davon gut

Political debate: mittel freien entwickelung liberalismus gemacht monarchischen indessen ganz bedeutung zeit glaubens weder regierung

Philosophy of science: geist denen idee lebens welt einzelnen vielmehr leben einzelnen philosophie schule recht staat partei

Education of the Jews: schon juden wenig sinne schule allgemeinen mag allerdings beziehung irgend sagen christlichen öffentlichen wenigstens

Education: bildung zeit erziehung schon humanismus schüler leben lehrer immer seyn kraft besonders deutsche ganz wohl allein aufgabe

Religion, Morality & Relationship to God: got zukunft humanität menschlichenleben wahre christenthum mensch freiheit demokratie wahrheit religiösen sagen gewalt welt politik

Study of languages: sprache sprachen zeit realschulen gesehen welt bildung amburg jugend schulen find gymnasien erfahrung habt werke neuen element

Notes

[1] See further Erling 2014: esp. 58–59, and Jacobi, Atteveldt & Welbers 2015.

[2] Blei & Lafferty 2006.

[3] The concept of *Humanismus* was coined in 1808 when Niethammer used it in his book *Der Streit des Philanthropinismus und Humanismus in der Theorie des Erziehungs-Unterrichts unsrer Zeit*. However, the tradition of German humanism dates back to the 15th and 16th centuries, when the ideas of Italian renaissance humanism spread across Europe. Accordingly, such concepts as *humanitas* and *studia humanitatis* are much older origin, dating back to antiquity.

[4] See further Bollenbeck 1994: 142–155.

[5] In addition to Georg Bollenbeck's book, the most extensive studies discussing 19th-century German humanism are by Landfester 1988 and van Bommel 2015.

[6] The book industry and the press were both growing in volume in the first part of the century, expanding even more dramatically as the 19th century neared its close. See further Erling & Tatlock 2014. Cf. St. Clair 2004.

[7] See further Brauer & Fridlund 2013: 159.

[8] Cf. Steinmetz & Freeden 2017: 2, 5.

[9] LDA (Latent Dirichlet allocation) was developed by David Blei and others in 2003 and MALLET (MAchine Learning for LanguagE Toolkit) was written by Andrew McCallum. For more information, see http://mallet.cs.umass.edu/about.php. See also *The Programming Historian* tutorial on MALLET, Graham, Weingart & Milligan 2012.

[10] The model provides as output three different files: topic 'state' assigning each word in the text to a topic, 'topic keys' consisting of the top words for each identified topic and the topic 'composition' consisting of allocation of percentage of every topic in each of the 95 files that were included in the analysis.

[11] DTM, Blei & Lafferty 2006.

[12] Hall, Jurafsly & Manning 2008: 364.

[13] Blei & Lafferty 2006.

[14] See ANNO webpage.

[15] See further, Hakkarainen 2020: 27–28.

[16] See further, e.g., Cordell 2015.

[17] Cf. Schonfield, Magnusson & Mimno 2017.

[18] See further Bollenbeck 1994: 142–155; van Bommel 2015.

[19] See, e.g., Anon. 1846; Anon. 1 & 4 August 1847; Anon. 1848.

[20] Bödeker 1982: 1123–1124.

[21] Ibid.:1121–1126. See also Hansson 1999: 77–106.

[22] See further Dussel 2011: 25–34; Stöber 2014: 141–142.

[23] Cf. Blevins 2010.

[24] Koselleck 1985. See also Steinmetz & Freeden 2017: 2, 5.

[25] However, even after filtering there was a minor increase in the percentage. In 1829, the proportional number for the word '*Zukunft*' was 0.010673; in 1850, it was 0.016537.

[26] In 1829, the proportional number for the word '*Wissenschaft*' was 0.01164-2238616473554; in 1850, it was 0.006096759758556382.

References

Anon. (1846, 7 January). Arnold Ruge und sein neuester Standpunkt. *Blätter für literarische Unterhaltung.*

Anon. (1847, 1 August). Arnold Ruge: Politischer Bilder aus der Zeit. *Blätter für literarische Unterhaltung.*

Anon. (1847, 4 August). Arnold Ruge: Politischer Bilder aus der Zeit. *Blätter für literarische Unterhaltung.*

Anon. (1848, 11 January). Polemische Briefe von Arnold Ruge: Reihe von 'Offenen Briefen zur Vertheidigung des Humanismus'. *Blätter für literarische Unterhaltung.*

ANNO – Austrian Newspapers Online (Austrian National Library). Retrieved from http://anno.onb.ac.at/faq.htm

Blei, D. M., & **Lafferty, J. D.** (2006, 25–29 June). *Dynamic Topic Models*. Paper presented at the ICML '06 Proceedings of the 23rd international conference on machine learning (pp. 113–120). Pittsburgh. Retrieved from https://mimno.infosci.cornell.edu/info6150/readings/dynamic_topic_models.pdf

Blevins, C. (2010). *Topic modeling Martha Ballard's diary*. Retrieved from https://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/

Bollenbeck, G. (1994). *Bildung und Kultur: Glanz und Elend eines deutschen Deutungsmusters*. Frankfurt am Main: Insel.

Brauer, R., & **Fridlund, M.** (2013). Historizing topic models: a distant reading of topic modeling texts within historical studies. In L. V. Nikiforova & N. V. Nikiforova (Eds.), *Cultural research in the context of 'digital humanities': proceedings of international conference 3–5 October 2013, St Petersburg* (pp. 152–163). St. Petersburg: Herzen State Pedagogical University and Publishing House Asterion. Retrieved from https://matsfridlund.files.wordpress.com/2014/04/publ2013brauerfridlundconf.pdf

Bödeker, H. E. (1982). Menschheit, Humanismus, Humanität. In O. Brunner, W. Conze & R. Koselleck (Eds.), *Geschichtliche Grundbegriffe: Historisches Lexikon zur politisch-sozialen Sprache in Deutschland*, Vol. 3: *H-Me* (pp. 1063–1128). Stuttgart: Klett-Cotta.

Cordell, R. (2015). Reprinting, circulation, and the network author in antebellum newspapers. *American Literary History, 3*(27), 417–445. DOI: https://doi.org/10.1093/alh/ajv028

Dussel, K. (2011). *Deutsche Tagespresse im 19. und 20. Jahrhundert*. 2nd edn. Berlin: Lit Verlag.

Erling, M. (2014). The location of literary history: topic modelling, network analysis, and the German novel, 1731–1864. In M. Erling & L. Tatlock (Eds.), *Distant readings: topologies of German culture in the long nineteenth century* (pp. 55–90). New York, NY: Camden House.

Erling, M., & **Tatlock, L.** (2014). Introduction: 'distant reading' and the historiography of nineteenth-century German literature. In M. Erling & L. Tatlock (Eds.), *Distant readings: topologies of German culture in the long nineteenth century* (pp. 1–25). New York, NY: Camden House.

Graham S., Weingart, S., & **Milligan, I.** (2012). Getting started with topic modeling and MALLET. *The Programming Historian* tutorial on MALLET. *The Programming Historian*, *1*. Retrieved from https://programming-historian.org/lessons/topic-modeling-and-mallet

Hakkarainen, H. (2020). Contagious humanism in early nineteenth-century German-language press. *Contributions to the History of Concepts 3*(15), 22–46. DOI: https://doi.org/10.3167/choc.2020.150102

Hall, D., Jurafsly, D., & **Manning, C. D.** (2008). Studying the history of ideas using topic models. In *EMNLP '08 Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 363–371). Honolulu, Hawaii. Retrieved from https://dl.acm.org/citation.cfm?id=1613763

Hansson, J. (1999). *Humanismens kris: Bildningsideal och kulturkritik i Sverige 1848–1933*. Stockholm: Brutus Östlings Bokförelag Symposion.

Jacobi, C., van Atteveldt, W., & **Welbers, K.** (2015). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism,* 1(4), 89–106. DOI: https://doi.org/10.1080/21670811.2015.1093271

Koselleck, R. (1985). *Futures past: on the semantics of historical time*. Translated by K. Tribe. Cambridge, MA: MIT Press.

Landfester, M. (1988). *Humanismus und Gesellschaft im 19. Jahrhundert*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Schonfield, A., Magnusson, M., & **Mimno, D.** (2017). 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*, Vol. 2: *Short papers*. Retrieved from http://aclweb.org/anthology/E17-2069

St. Clair, W. (2004). *The reading nation in the Romantic period*. Cambridge: Cambridge University Press.

Steinmetz, W., & **Freeden, M.** (2017). Introduction: conceptual history. In W. Steinmetz, M. Freeden & J. Fernández-Sebastián (Eds.), *Conceptual history in the European space* (pp. 1–46). New York, NY: Berghahn Books.

Stöber, R. (2014). *Deutsche Pressegeschichte*. 3rd edn. Munich: UVK Verlagsgesellschaft Konstanz.

van Bommel, B. (2015). *Classical humanism and the challenge of modernity: debates on classical education in 19th century Germany*. Berlin: De Gruyter.