Topical Discourse Structures

Using Topic Modeling in Discourse Analysis Approaches

Fabian Brinkmann, Ruhr-Universität Bochum

With the amount textual data available to researchers rapidly increasing, the Humanities and Social Sciences have to deal with new challenges in utilizing these large quantities of texts. For several decades, Digital Humanities have offered a multitude of tools for computer-assisted or -driven research. This article will explore how distant reading approaches in general and topic modeling in particular can be utilized in discourse analysis. It will present theories and methods that work well together and can be applied to different research projects, using a combination of Structural Topic Modeling, developed by Roberts, Stewart and Tingley (2018) and Siegfried Jäger's Critical Discourse Analysis approach as an example.

Keywords: distant reading; textual statistics; topic modeling; discourse; critical discourse analysis

Contemporary scholars have access to vast amounts of textual data ranging from corpora of governmental reports to digital newspaper archives and the incredible amount of text written daily on social media like Twitter. These amounts of content pose new and important challenges to researchers (Lucas *et al*. 2015, 2), that in most cases are not yet completely understood by the research community, since the grand scale of this data is a relatively new development. In this context, some estimates even claim that 90% of the data existing in 2012 did not exist two years

Brinkmann, Fabian "Topical Discourse Structures: Using topic modeling in discourse analysis Approaches" *HUMAN IT* 14.3 (2019): 83–114.

before that (Sharma *et al*. 2014, 139). While organizations like JSTOR offer access to huge digital libraries, a simple keyword search of such a collection of documents is often not enough for more specific research questions. This is due to the fact that the researcher can only read a limited amount of text in a close reading approach and such an approach prevents a more structured search through document corpora ranging up towards tens or even hundreds of thousands of documents due to the sheer amount of words to be read (Blei & Lafferty 2009, 71). Even if just a relatively small amount of the available textual data is relevant to researchers, different corpora need specific approaches to identify and deal with issues of interest in different languages and different topics (Lucas *et al*. 2015, 21).

As such, the question posed by David Mimno, "How, if at all, should the work of humanistic scholarship adapt to the presence of orders of magnitude more potential source material?" (Mimno 2012a, 1), is still vital for the future of the humanities. Clearly, the abilities of computers in text mining, information retrieval and statistical analysis offer chances to complement traditional humanistic scholarship. At the same time, computer programs make it possible to ask new and extended research questions and gain additional insights into discourse and human behavior on a much larger, collective scale (Lazer *et al*. 2009, 722). With a broader foundation of data, insights that could previously only be approximated or deducted from case examples can come into the focus of modern research (King 2009, 92). These new ways to deal with large amounts of data have already been implemented in fields like biology and physics. However, the implementation of computer-driven approaches has been much slower in the humanities and social sciences (Lazer *et al*. 2009, 721).

The main problem with traditional approaches of close reading, especially in the study of political, social or historical discourse, is their unscalability. The amount of textual data available cannot possibly be explored by a single researcher and even if it were possible, important

details would be lost, since the researcher's attention could only be directed at a small amount of the available texts (Chuang *et al*. 2015, 176). As an alternative, literary scholar Franco Moretti has developed the concept of distant reading, stating: "It allows you to focus on units that are much smaller or much larger than the text: devices, themes, tropes—or genres and systems" (Moretti 2002, 57-58). Others have coined the term "surface reading" in contrast to "symptomatic reading", which focuses only on specific texts or examples within a much larger corpus (Erlin & Tatlock 2014, 2). Especially, the study of discourse often stays in the realm of symptomatic reading, looking at example texts and case studies to reach conclusions about much larger issues.

Computational text analysis first started in 1949 with Italian Jesuit priest Roberto Busa's Index Thomisticus, an index verborum of Thomas Aquinas' writings. The first volume was published in 1974 (Busa 1980). While prior to the 1960s humanities computing was focusing only on the sentence layer of texts, first approaches towards the text as a complete linguistic entity became more common in the 60s (Stede 2007, 19). By now, automated text analysis is applied in a multitude of research fields including Geo- und Sociolinguistics (*e.g*. Nerbonne 2009; Eisenstein *et al*. 2010; O'Connor *et al*. 2010), Political Science (*e.g*. King & Lowe 2003; Shellman 2008; Grimmer 2010; Quinn *et al*. 2010; Black *et al*. 2011; Metaxas, Eni & Gayo-Avello 2011; Stephens-Davidowitz & Seth 2012), History and Historical Linguistics (*e.g*. Horton *et al*. 2009; Bamman & Crane 2011), Social Psychology (*e.g*. Tausczik, Yla & Pennebaker 2009; Golder & Macy 2011), Economics, Finance and Management (*e.g*. Das, Sanjiv & Chen 2007; Tetlock 2007; Askitas & Zimmermann 2009; Joshi *et al*. 2010; Loughran & McDonald 2011), and Literary Studies (*e.g*. Holmes 1998; Argamon *et al*. 2009; Craig & Kinney 2009).

However, modern computer science has developed models that go far beyond simple exploratory or text mining approaches. Latent variable models are able to even uncover "interpretable, low-dimensional subspaces" (Mimno *et al*. 2011, 262). These statistical models are able to

represent the complexity of large document collections based on their topical structures (Mcaufliffe & Blei 2008, 121). One of the most common approaches in this field is called topic modeling.

While the ultimate potential of topic modeling is still being explored (Daoud & Kohl 2016, 7), it has already been applied to a number of different research questions. These include an analysis of all editions of Science from 1990 to 1999 (Blei & Lafferty 2007), the prediction of U.S. Supreme Court decisions (Mimno 2012b, 84), topical structures in New York Times articles (Ibid., 88-89) issues of the State of the Union addresses (Ibid., 91-93), psychological accounts of linguistic processing and semantic memory (Griffiths, Steyvers & Tenenbaum 2007, 237), scholarly discourse in Literary Studies (Goldstone & Underwood 2014), the influence of Darwin on Nordic writers (Tangherlini & Leonard 2013, 735-736) folkloric topics in Nordic literature (Ibid., 742), data mining in JSTOR articles (Mimno 2012a), ideology in Economics papers (Jelveh, Kogut & Naidu 2014), structures of political information in Russia (Baturo & Mikhaylov 2013), anti-Americanism on Arabic Twitter (Jamal *et al*. 2015), opposition in British House of Commons debates (Eggers & Spirling 2014), censorship in China (King, Pan & Roberts 2013), U.S. national security strategies (Mohr *et al*. 2013), scholarly correspondences in the 17th century (Wittek & Ravenek 2011), and the recommendation engine of the New York Times, which is based on a topic modeling algorithm (Spangher 2015).

Looking at the different research areas, it should become clear that a topic modeling approach can be adequately used in political, social and historical research, as has been pointed out by Andrew Goldstone and Ted Underwood (2014), Jonathan Slapin and Sven-Oliver Proksch (2009), Amr Ahmed and Eric Xing (2010), Bonnie Webber and Joshi Aravind (2012), Justin Grimmer and Brandon Stewart (2013), as well as Paul DiMaggio, Manish Nag and David Blei (2013).

To expand on this, this paper will review some papers in which topic

modeling has been used to study discourse in order to present a theoretical combination of the approaches of topic modeling and discourse analysis. First, it will lay out the technical and statistical foundation of topic modeling. This will be exemplified with the Structural Topic Model (STM), developed by Molly Roberts, Brandon Stewart and Dustin Tingley, which offers deeper possibilities for the discourse analyst. Furthermore, it will suggest that specific aspects of discourse can be investigated with the aid of statistical topic modeling, mainly drawing from the concept of discourse strands in Siegfried Jäger's approach to CDA.

Topic Modeling As a Tool for Text Analysis

With the increasing need to analyze large textual corpora, the application of hierarchical statistical models on topics has gathered growing interest in recent years. Latent variable models assume that a set of complex data indeed shows simpler patterns not visible at first glance (Blei 2014, 204). However, these underlying structures of topics or ideas are not easily discovered by a close reading approach, especially when dealing with large corpora. Thus, automated methods for exploring and browsing these document structures have to be implemented (Blei & Lafferty 2009, 71).

Topic models are "probabilistic models for uncovering the underlying semantic structure of a document collection based on a hierarchical Bayesian analysis of the original texts" based on Latent Dirichlet Allocation (LDA) (Ibid., 71). They assume that the hidden structure of a corpus can be observed in the distribution of topics across the documents (Blei 2012, 79). In that sense, they mainly uncover patterns of word use across a corpus of documents without having any prior knowledge of these discovered semantic structures (Blei & Lafferty 2009, 72). It should be noted, that topic modeling as a term was retroactively applied, although it fits the intuitive assumption that documents exhibit a number of different topics (Blei 2012, 78). In simple technical terms a topic model "uses a small number of distributions over a vocabulary to

describe a document collection" (Lafferty & Blei 2006, 148); the inferred semantic distributions can then be organized and analyzed by the researcher as corresponding to the underlying topics in the corpus (Mimno 2011, 1).

Topics are defined as distinct semantically associated co-occurrences of words (Wallach, Mimno & McCallum 2009, 1981). They represent topical structures of content, are presented in relative frequency to the whole topical structure of the corpus, and can be helpful in observing the underlying semantic texture of large textual corpora, which are usually unstructured, especially with regards to topical content (Roberts, Stewart & Tingley 2016). It is important to note that the model is based on the assumption that the topics exist before the text and topics themselves generate the content of documents (Blei 2012, 78). Thus, topic modeling, being an unsupervised method, infers the content of the topics from the textual data without any prior input by the researcher (Roberts *et al.* 2014, 1066). This does not mean that no prior choices have to be made. Assumptions like the number of generated topics are an important part of the process and have to be made explicit as the method parameters, which can vary for to different research questions and/or corpora (Goldstone & Underwood 2014).

In its mathematical sense, LDA, the most common topic modeling algorithm, defines topics as "a distribution over a fixed vocabulary of terms" (Blei & Lafferty 2009, 72). Usually each document is treated as a vector of word frequencies, meaning word order is ignored. This is called the 'bag-of-words' approach, where K topics with a multinomial distribution over V words are associated with a corpus of documents and the individual documents exhibit these topics with different proportions (Blei & Laffery 2009, 72-73). A corpus of newspaper articles, for example, might reveal topic vectors interpretable as politics, sports, culture, crime, etc., and a particular article could be described as exhibiting these topics in different proportions (for example the topics of politics and

crime in an article about government corruption). Through the application of mixed-membership models each topic can be assigned to multiple documents and each document can exhibit multiple topics (Blei 2014, 211-212).

Aside from extracting topical structures from textual corpora, topic models can also be used for a number of further tasks including distributional semantics, word sense induction, information retrieval, classification, prediction and collaborative filtering (Mimno 2011; Stevens *et al*. 2012).

Topic models can function as a useful approach to text analysis in the social sciences and humanities in general as well as discourse analysis in particular, because they usually offer understandable and feasible readings of corpora and texts (Mohr & Bogdanov 2013, 546). The main strengths of this approach are the abandonment of prior annotations for an unsupervised model (Hu *et al*. 2014, 424), the resulting possibility to study much larger corpora (Genovese 2015, 4), the possibility to study research issues "through a macroscopic lens" (Mohr & Bogdanov 2013, 561) and the ability for each document to be linked with multiple topics in order to represent the thematic combinations in different documents (Rosen-Zvi *et al*. 2010, 3-4). In a way, topic modeling allows the researcher to "work backward" (Erlin 2014, 59). If texts are assembled from prior topical contents, the algorithm allows these contents of the original topical distribution that informed the investigated corpus to become visible (Graham, Milligan & Weingart 2016, 113).

Topic modeling has, however, not remained without criticism from humanities researchers. Indeed, a poorly implemented topic modeling approach will at best be unsurprising or at worst unhelpful if not plain uninterpretable (Schmidt 2012), and all methods of automatic content analysis should be validated by close reading and comparisons of competing models (Grimmer & Stewart 2013). Lastly, one gap remains in most of the 'classical' approaches towards topic modeling. The models do not

include document metadata and thus are unable to represent the distributions of topics across different covariates like time, actors or genre (Mimno 2012b, 4).

Expanding on Topic Models: Structural Topic Modeling

Usually text documents come with several instances of additional data like authors, titles, dates or subjects. This metadata can be used to make further inferences about the documents. In the case of a corpus, one has to distinguish between two different types of metadata: document metadata and collection metadata. While the former refers to data that is specific to a document, like classification into genre, author or date of publication, the latter signifies data that describe the corpus as a whole and is not specific to single documents, like the date of the corpus creation (Feinerer, Hornik & Meyer 2008, 8).

Especially, document metadata is of interest for topic modeling approaches, and extensions to topic modeling packages that take into account the accompanying metadata in order to discover patterns associated to the metadata have been a staple of topic modeling development for years (Mimno & McCallum 2008). The goal of these extensions of topic modeling is the discovery of associations between text and metadata and the identification of underlying patterns of text collections (Mimno 2012b, 98).

Two of the more basic models to investigate corpora with the use of the metadata are the author-topic model and topics-over-time. The author-topic model, for example the one developed by Rosen-Zvi, Steyvers and Smyth (2004), assigns topics to specific authors and thus can be used to analyze the topical foci of text creators and offers conclusions about the connections between authors and topics (Blei 2012, 83). The topics-over-time algorithm associates topics with distributions over timestamps and is able to represent the rise and fall of topics in a diachronic perspective (Wang & McCallum 2006, 424-425). A more general application

can found in the Dynamic Topic Model. While common LDA assumes that documents are generally interchangeable, this is not the case when using metadata. The Dynamic Topic Model slices the documents based on metadata (*e.g.* year) and models topic proportions according to these slices (Blei & Lafferty 2009, 84).

The Structural Topic Model (STM), a package in R developed by Molly Roberts, Brandon Stewart and Dustin Tingley and an extension of the models described above, offers a more adjustable way to plot the relationship between topics and metadata. Possible associations with metadata include the place of text origin, the author, characteristics of the author, date of origin as well as genre. At the same time, STM can estimate correlations between topics (*i.e.* what topics are closely associated with one another) and can create graphical depictions of these correlations. It offers "a range of features from model selection to extensive plotting and visualization options" (Roberts, Stewart & Tingley 2018, 1). Because of the general applicability of STM, the amount of papers using the model and the number of different functions included in the package, the STM shall serve as an exemplary inclusion of document metadata in a topic model for the sake of the arguments presented in this paper.

STM was specifically developed with the idea in mind that the contents of texts (or the underlying discourse) changes over time; a fact that common LDA applications usually ignore and thus cannot account for changes in language or terminology (Riddell 2014, 108). Based on LDA and the Correlated Topic Model, STM set out to offer a flexible way to analyze the relationship of metadata to texts and topics (Roberts, Stewart & Tingley 2018, 1). Applications include the analysis of Arabic fatwas and Arabic as well as Chinese social media responses to the events surrounding Edward Snowden (Lucas *et al.* 2015), American political blogs during the 2008 presidential election (Roberts *et al.* 2016), debates about organ donation on Facebook (Bail 2016), politics in papal encyclicals

(Genovese 2015), ideological discussions about climate change (Farrell 2016), as well as the uncovering of constitutional archetypes (Law 2016).

The two main points of analysis for the STM are topical prevalence or topical content (Roberts *et al.* 2013, 1-2). Topical prevalence represents the frequency with that a topic is discussed (*i.e.* was a topic more prevalent in the parliamentary debates of a specific year or does a newspaper write more about a specific topic than another), while topical content allows for the observation of how a particular topic is discussed (*i.e.* does a superior talk differently to his customers than his employees or does a newspaper write differently about a topic than another) (Roberts, Stewart & Tingley 2018, 7). The inclusion of these analysis aspects allows for an encompassing analysis of political, social and/or historical discourse, which is important, since researchers are more interested in the way observable covariates affect the content of a document or corpus than what the corpus is generally about. STM offers a flexible way of plotting these relationships that scales well with different sizes of document collections (Roberts, Stewart & Airoldi 2016, 2).

The Analysis of Discourse

How then can the approach of STM be applied in order to undertake a structured analysis of discourse? To answer this question, one must first discuss what discourse actually means. This paper will, on a fundamental level, work with a basic definition of discourse, based on the assumption that rules exist, which inform what can and what cannot be said in a specific circumstance. Taking into account that rules about what can be said give order to social structure, this approach also sees discourse as the basis for what can be thought and what can be done, which puts discourse in the center of all analyses involving human acts (Landwehr 2006, 107-108). If this is true, discourse analysis provides a useful tool for answering questions of humanities research, such as history, cultural studies and sociology, because these disciplines all involve human

behavior and language and aim to answer questions about social relations, processes of identification as well as social and political decision-making (Johnstone 2002, 7). As such, most approaches to discourse analysis "aim to provide a better understanding of socio-cultural aspects of texts, via socially situated accounts of texts" (Kress 1990, 84). Especially in the last decades, this approach to discourse analysis, which views discourse as a social and historical phenomenon and looks at the rules that produce social and historical knowledge, became increasingly popular (Landwehr 2006, 111).

The foundation for these assumptions lies in the linguistic turn of the 1960s, where the epistemological position that language plays an important role not only for the understanding of realities but also, through conventions of characterization and language, for the constitution how realities itself is perceived was first popularized (Sarasin 2003, 11-12).

Discourse analysis is a method to investigate texts and their communicative foundation. Since texts usually have a discursive function in the shaping of social, cultural and historical realities it is an important part of discourse analysis to present this function in a stringent fashion and achieve insights about the political and social applications of discourse (Chimombo & Roseberry 1998, ix-x). While most discourse analysts cite the discourse notions of scholars like Foucault, Habermas, Laclau or Luhmann (Wodak & Meyer 2009, 2-3), their theories often become a backdrop for more practical applications of discourse analysis. Instead of looking at discourse as one overarching set of rules that governs society, these applications use discourse as a count noun and look at how sets of discourses and ideas within a specific field "influence [each other] and are influenced" by each other, creating "conventionalized sets" of language within the field, which can be analyzed for a deeper understanding of how an issue is constructed and discussed through statements (Parker 1992, 5; Johnstone 2002, 3).

The methods and theories presented in this paper focus on a particular style of Discourse Analysis: Critical Discourse Analysis (CDA). CDA is based on the general assumption that discourses do not just represent social realities, but instead have a life on their own, because the discursive actors change and shape social realities through their discursive acts and statements (Jäger 2012, 33-35). Thus, "CDA sees discourse – language use in speech and writing – as a form of social practice" (Fairclough & Wodak 1997, 258). The goal of CDA is determining the "atoms of discourse", that define the conventionalized way to talk about an issue, meaning: what can be said and thought in specific circumstances (Jäger 2012, 8-12). CDA was first developed by a group of scholars in the early 1990s (Wodak & Meyer 2009, 3). Since then, several branches of CDA have been developed. Among the most famous are Norman Fairclough's Dialectical-Relational Approach, the Theo van Leeuwen's Social Actors Approach, the Ruth Wodak's Discourse-Historical Approach, Teun van Dijk's Sociocognitive Approach as well as the CDA branch of the Duisburg School led by Siegfried Jäger, which will form the foundation of the theoretical developments of this paper.

This branch of CDA utilizes a set of terms, which form a consistent theory of discourse. First, there are 'fragments of discourse' which are thematically homogenous parts of texts (Jäger & Zimmermann 2010, 39). Usually a text consists of several fragments of discourse; this is called 'discourse entanglement' (Jäger 2012, 87). On the next layer one can find 'discourse strands', defined as thematically consistent trends of discourse, which regularly appear in an overall societal discourse. These topics/themes usually exhibit a number of smaller topics themselves and/or are composed of different fragments discourse. Such formations of discourse can be used in both synchronic and diachronic analyses of the underlying discursive structures (Ibid., 80-81). Other discourse analysts have utilized similar concepts, using the term 'topos' to describe patterns of argumentation that come up in similar ways in different texts

and are connected in their goal of construing similar issues (Bernard 2009, 32). These thematic strands are often defined by the 'discursive position' of the speaking subject, person, group or medium. The discursive position can be seen as the space from which utterances in discourse are made, which itself is shaped by the diverse discourse in which the subject is entangled (Jäger 1996, 47). This is followed by 'layers of discourse', the social realities from which subjects engage with the overall discourse. They inform discursive positions as well as discourse strands (Jäger & Zimmermann 2010, 38).

Jäger also utilizes the terms 'discourse community', groups which have relatively homogenous beliefs, ideologies or 'discursive positions', 'discursive knots', entanglements between different discourse strands, as well as 'discursive context', which is important to understand the diachronic and synchronic qualities of a specific discourse strand (Ibid., 40-42). Other approaches have also used the term 'leading words' (German: 'Leitvokabeln'), which are terms that represent specific aspects of discourse, play an important role within the discursive structure and show relationships within a pattern of entanglement (Busse & Teubert 1994, 22).

An important aspect of CDA can also be found in the "intertextual and interdiscursive relationships between utterances, texts, genres and discourses" (Bhatia 2006, 178). Intertextuality describes the linkings between different texts, both in a synchronic and diachronic measure, and is used to analyze the transfer of arguments, topoi and topics between texts (Ibid.). Interdiscursivity, on the other hand, means that discourses are connected with each other and that a discourse on a specific topic can also refer to topics from other discourses (Reisigl & Wodak 2009, 90). In fact, according to CDA, all texts are multidiscursive, since they can refer to a range of topics, discourses and fields of knowledge (Mogashoa 2014, 108).

All of this is especially important, since discourses are closely entangled with each other. Jäger calls this discursive "Gewimmel" (loosely

translated as 'milling mass'). Untangling this mass is the task of discourse analysis (Jäger & Zimmermann 2010, 15-16). The concepts outlined above can be used to structure this milling mass. However, it should be noted that the applications of all of these concepts for discourse analysis function as specific tools in CDA. Jäger himself has defined CDA as a 'tool box', from which concepts and approaches can be plucked according to the analytic needs, meaning that not all of these concepts have to be utilized in every analysis (Jäger 2012, 8). As for all analyses of social communication, even though the goal is the analysis of how social truths are constructed, hermeneutic limitations apply. Thus, it should also be noted that, "THE RIGHT interpretation does not exist; a hermeneutic approach is necessary. Interpretations can be more or less plausible or adequate, but they cannot be true" (Wodak & Ludwig 1999, 13).

Topic Modeling and Discourse Analysis

According to Siegfried Jäger, the general goal of discourse analysis is the analysis of a discourse strand or several entangled discourse strands in diachronic and synchronic perspective. He points out that discourse strands are put together from several discourse fragments, which he likens to Foucault's utterances in the Archaeology of Knowledge and which signify different topics or themes (Jäger 1999, 136-137). To analyze these topics it is important to note that discourses in their grand sense usually cannot be found. What the researcher is working with are rather pieces of discourse (Parker 1992, 6). While such individual pieces are surely informative, they are not as meaningful as a grand picture of the structures of discourse. In such a case, statistical topic modeling can be used to identify clustered groups of words that signify overall semantic structures in discourse (Mimno 2012a, 6).

In discourse analysis grand-scale analyses of texts have long been an important research field. As CDA analyst Teun van Dijk points out: "I often advocate beginning Critical Analysis with an analysis of semantic

macrostructures, that is with a study of global meanings, topics or themes. These are what discourses are (globally) about" (van Dijk 2009, 68). While not utilizing computational approaches, he still stresses that topics or themes represent significant textual macrostructures, because they influence other discursive structures such as the global coherence of discourse and have the most effect on the discourse participants.

Another discourse analyst, Achim Landwehr, emphasizes the importance of identifying the narrative patterns or the macrostructures of a text corpus in a first step of discourse analysis. For Landwehr the most important signifier of these macrostructures are the topics of a text (Landwehr 2001, 114-115), because regularly repeated statements about a topic can be used to ascertain the focal points of how the topic is discussed (Keller 2006, 54). Although also not working with digital methods, he points out the possibility of a quantitative approach to identify these patterns (Landwehr 2001, 116). CDA analyst Norman Fairclough also stresses that what is 'experienced' from a text or utterance is usually referred to as the topic (Fairclough 2007, 133).

If the researchers view topics as "semantic macropositions", discourse can generally be about any topic, although it may show preferred topics that are expressed as pivotal discourse points or strands. These topics will also allude to other topics, since discourses are usually not only about a single topic. For example an argument about immigration policies will often not only deal with politics, but also with issues of minorities (van Dijk 1997, 25-26). Both the general topic proportions of topic modeling (*e.g.* which topics are discussed the most) and the term proportions within topics (*e.g.* which words are used the most within a specific topic) can be used to analyze these topical structures or semantic macropositions. Following the positions outlined above, a computational analysis of the thematic structures in a corpus can also be helpful to analyze ideologies or shared sets of beliefs (Ahmed & Xing 2010, 1140).

This possibility is of great value for discourse analysts of every research field. Not only does it offer the possibility to analyze of much larger text corpora, but it also eliminates the error-prone need for humans in the process of indexing and prevents the researchers from applying "their own preconceived identification of topics" (Newman & Block 2006, 766). While topic modeling is not infallible and certainly cannot fully explain social causation, it helps to avoid simple causal explanations by offering a larger context and creating a more nuanced account across multiple facets of the issue at hand. In a best-case scenario, topic modeling is even capable of breaking apart preconceived positions by revealing new and exciting topical structures latent in the texts (Goldstone & Underwood 2014).

However, it should be noted that the topic is not completely congruent with the whole materiality of discourse. Topic Modeling can offer valuable insights into topical structures and patterns, but it is not capable of making deep linguistic analyses or identifying blank spaces in discourse (Haslinger 2006, 41). Like Critical Discourse Analysis, computational analysis of text corpora should indeed be treated as a tool box which is valuable for some research questions, but cannot be universally applied to all research and is certainly not a silver bullet for answering complicated research questions.

Similarly, the choice of model is important, because different models offer different approaches and different results (Blei & Lafferty 2009, 82).

When an author describes events, entities and ideas, they are expressed as topics in the sense of topic modeling (Mimno 2012b, 4). Thus, a topic model is able to assist in analyses of semantic representations of content (Griffiths, Steyvers & Tenenbaum 2007, 212). Since the analysis of discourse strands is, according to Siegfried Jäger, mostly about analyzing thematically consistent trends of discourse, which regularly appear in an overall societal discourse, it seems likely that, based on the

theoretical assumptions of statistical modeling, topic models can provide useful insights into the structures and entanglements of discourse strands. While this will typically be done through the analysis of central words and statements, the topics of topic modeling also show co-occurring words as context which can be used to improve the analytical approach and resolve ambiguity (Marshall 2013, 708). Likewise, topic modeling is also able to assess latent positions and arguments which underlie discourse but are difficult to grasp in a close reading approach (Slapin & Proksch 2008, 706). Since discourse strands are primarily defined through their thematic component, they show a close resemblance to the theoretical description of the topics of topic modeling. If that is the case, topic modeling is able to show the consistency (or inconsistency) of these thematic strands and offers insights into the internal semantic structures of discourse strands. Jäger points out that the analysis of thematic discourse strands is used to show the general contents and arguments of a discourse as a common denominator with regards to content (Jäger & Zimmermann 2010, 29-30). Since topic modeling was primarily developed to offer insights into thematic patterns, which suggest that it could complement a CDA approach in a valuable way.

Models that include document metadata, such as those that can be created by the 'stm' package offer additional insights into the structure of a corpus. It is almost a platitude that issues, arguments and topics rise and fall in history (Griffiths & Steyvers 2004, 5232), often in accordance with specific (discursive) events. Understanding these diachronic dynamics is made possible through the analysis of topical prevalence across a temporal covariate offered in topic modeling. Describing the changes of topical prevalence over time thus can offer insights into the diachronic layer of discourse strands, since discourse strands appear in accordance with discursive events and longer periods of time have to be included in the analysis of discourse strands, "in order to identify the changes, ruptures, ebbing and recurrences of a discourse strand" (Jäger & Maier

2009, 51). This means that topic modeling can be used to uncover chronological trends in topical structures (Goldstone & Underwood 2014). Spikes in topic proportions are interesting research results, which, with sufficient contextual knowledge, can yield insights into the influence of discursive events on overall discourse (Miller 2013, 643). Additionally, the topical content approach can also yield results about shifts in vocabulary use when dealing with a specific topic in different time periods (Hall, Jurafsky & Manning 2008, 363), offering additional insights into the diachronic changes of a discourse strand.

With respect to synchronic analysis, an actor-centered approach can offer interesting results. Researching "the trajectories of individual authors across […] topics" (Anderson, McFarland & Jurafsky 2012, 13) is made possible by using an actor covariate in a topic model. This can lead to insights into thematic foci of different actors (topical prevalence) and different use of vocabulary (topical content). The latter can be used to investigate contention between actors in specific discourse strands, because ideology is often expressed through the use of different terms (Schäffner 1996, 2). Combining this with a diachronic analysis can lead to results about turning points in discursive contention (Anderson, McFarland & Jurafsky 2012, 13).

Intertextuality in CDA means that "texts are linked to other texts, both in the past and in the present" (Reisigl & Wodak 2009, 90). According to CDA, these connections are represented in a multitude of ways. While actors are certainly important for the connectivity between different texts, the focal point of intertextuality is the reference to a topic or an argument and may thus be able to be expressed by a topic modeling approach since different texts are weighed by topic proportions in the model. This could also yield insights into interdiscursivity by looking at the different topics appearing together across several documents, because the links between discourses are "primarily topic-related" (Reisigl & Wodak 2009, 90). On a qualitative layer, this could be verified by the

production of example texts with high proportions of a specific topic by the model (Roberts, Stewart & Tingley n.d., 10-11).

Looking at the possibilities of Structural Topic Modeling, they suggest that those can be used to identify and analyze discourse strands, which are "made up of discourse fragments of the same topics" (Jäger 2012, 80), albeit in a very different scope than the qualitative and close approach by Siegfried Jäger.

The Importance of Close Reading

Topic modeling certainly offers a new and promising approach to text analysis. However, it is still starting out and a lot of approaches are still in development (Jacobi, van Atteveldt & Welbers 2016, 89). Critics have pointed out that topic modeling fails at including syntax and context into its algorithms, which sometimes can ignore the ambiguity of different sentiments in a purely vocabulary-based approach (Slapin & Proksch 2009, 324), although the general topic is, of course, still extracted by the algorithm.

Likewise, criticism has been leveled at the quality of topic modeling results. Indeed, topic modeling can sometimes yield mixed results. Topics are usually considered 'good', if their terms can be understood as semantically coherent, which might not be the case for every topic (Mimno & Blei 2011, 262-264). This is especially the case with topics comprised of generic terms. However, two answers can be given towards this criticism. The first is expressed in David Blei's crucial question: "Is my model good enough in the ways that matter?" (Blei 2014, 226). Even if a model expresses some uninterpretable topics, it can still be of use in answering the research question and the 'good' topics can still adequately express the topical structure of a corpus.

The future of topic modeling also "lies in close collaborations between domain experts and modelers" (Ibid., 205) and the automated analysis of topic modeling does not free the researchers from their own

interpretative effort and a hermeneutic methodology (Goldstone & Underwood 2014). Indeed, results of quantitative text analysis will always need to be verified in a qualitative approach. This is the case for both the internal coherence of the topics and their congruence with the modeled texts and no topic modeling approach will ever replace a careful qualitative reading of texts. "Rather, the methods that we profile here are best thought of as amplifying and augmenting careful reading and thoughtful analysis." (Grimmer & Stewart 2013, 268). Close and distant reading should not be seen as opposing ways of analysis (Mimno 2012a, 17). The challenge for topic modeling users is how to best identify the optimal way to combine human close reading and automated distant reading in gainful analysis (Grimmer & Stewart 2013, 270).

Good contextual knowledge will both be needed for a sensible interpretation of the proportional topics list as well as for understanding the generated lists of words in the context of the research issue at hand (Miller 2013, 644). Without contextual knowledge, topic modeling results will often be uninterpretable. Likewise, a close reading of example texts to further look at the structures of the discourse in question will be unavoidable. In this way, a qualitative discourse analysis of the investigated discourse strand becomes possible and offers opportunities to perform 'classical' (Critical) Discourse Analysis to complement the computational approach of topic modeling.

Conclusion

While certainly only the surface of the vast possibilities topic modeling offers has been scratched so far (Blei 2014, 218), it already offers a useful tool that "can help us grapple with the subtle interpretive problems endemic to cultural history, where a change is often determined by multiple causes" (Goldstone & Underwood 2014). Even though the quantification of research has some scholars worrying about interpretability, one can remain optimistic about the future of automated content

analysis in general and topic modeling in particular. With more and more textual collections available digitally, large scale analysis of corpora will certainly become increasingly important in the future (Cohen & Rosenzweig 2006, 80).

This paper reviewed some of the literature of topic modeling to suggest that topic modeling offers new and unique ways to analyze the thematic structures of discourse, called discourse strands in Siegfried Jäger's approach to CDA. With discourse strands being thematically coherent threads of discourse within overall discourse that change across time and are utilized differently by different actors in different discursive position, Structural Topic Modeling in particular may very well be capable of uncovering synchronic and diachronic changes in topical discourse structures and can thus complement the approach of CDA in exciting ways.

Nevertheless, algorithmic approaches cannot completely substitute close reading of texts. They can offer new insights and more possibilities for corpus exploration, but the researcher will always have to utilize prominent example texts and vast contextual knowledge to adequately grasp the structures of thematic discourse across different actors and time periods.

Fabian Brinkmann is a Ph.D. student in History at the Centre for Mediterranean Studies at Ruhr-University Bochum, Germany. His project is titled: 'A computer-assisted analysis of Turkish foreign policy towards Sub-Saharan Africa: Diachronic topics and networks (2002-2016)". His research interests include contemporary Turkish history, Turkish foreign policy, digital humanities, text statistics.

Contact: fabian.brinkmann@rub.de

References

AHMED, AMR & ERIC XING (2010). "Staying Informed: Supervised and Semi-Supervised Multi-view Topical Analysis of Ideological Perspective." *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. 1140–1150.

ANDERSON, ASHTON, DAN MCFARLAND & DAN JURAFSKY (2012). "Towards a Computational History of the ACL: 1980-2008." *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*. 13-–21.

ARGAMON, SHLOMO *ET AL*. (2009). "Gender, Race, and Nationality in Black Drama, 1950-2006: Mining Differences in Language Use in Authors and Their Characters." *Digital Humanities Quarterly* 3.2. <http://www.digitalhumanities.org/dhq/vol/3/2/000043/000043.html> [2019-12-20]

ASKITAS, NIKOLAOS & KLAUS F. ZIMMERMANN (2009). "Google Econometrics and Unemployment Forecasting." *Applied Economic Quarterly* 55.2: 107–120.

BAIL, CHRISTOPHER A. (2016). "Cultural Carrying Capacity: Organ Donation Advocacy, Discursive Framing, and Social Media Engagement." *Social Science & Medicine* 165: 280–288.

BAMMAN, DAVID & GREGORY CRANE (2011). "Measuring Historical Word Sense Variation." *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*. 1–10.

BATURO, ALEXANDER & SLAVA MIKHAYLOV (2013). "Life of Brian Revisited: Assessing Informational and Non-informational Leadership Tools." *Political Science Research and Methods* 1.1:139–157.

BERNARD, TARYN (2009). *Justificatory Discourse of the Perpetrator in TRC Testimonies: A Discourse-historical Analysis*. Stellenbosch University. <https://core.ac.uk/reader/37319911> [2019-12-20]

ADITI BHATIA, ADITI (2006). "Critical Discourse Analysis of Political Press Conferences." *Discourse & Society*, 17.2: 173–203.

BLACK, RYAN C. *ET AL*. (2011). "Emotions, Oral Arguments and Supreme Court Decision Making." *The Journal of Politics* 73.2: 572–581.

BLEI, DAVID M. (2012). "Probabilistic Topic Models." *Communications of the ACM* 55.4: 77–84.

BLEI, DAVID M. (2014). "Build, Compute, Critique, Repeat: Data Analysis with Latent Variable Models." *Annual Review of Statistics and its Application* 1; 203–232.

BLEI, DAVID M., & JOHN D. LAFFERTY (2007). "A Correlated Topic Model of Science." *The Annals of Applied Statistics* 1.1: 17–35.

BLEI, DAVID M. & JOHN D. LAFFERTY (2009). "Topic Models." *Text Mining: Classification, Clustering, and Applications*. Eds. Ashok N. Srivastava & Mehran Sahami. Boca Raton: CRC Press. 71–94.

BUSA, ROBERTO (1980). "The Annals of Humanities Computing. The Index Thomisticus." *Computers and the Humanities* 14.2: 443–459.

BUSSE, DIETRICH & WOLFGANG TEUBERT (1994). "Ist Diskurs ein Sprachwissenschaftliches Objekt?: Zur Methodenfrage der Historischen Semantik." *Begriffsgeschichte und Diskursgeschichte: Methodenfragen und Forschungsergebnisse der Historische Semantik*. Eds. Dietrich Busse, Fritz Hermanns, & Wolfgang Teubert, Opladen: Westdeutscher Verlag. 10–28.

CHIMOMBO, MOIRA & ROBERT L. ROSEBERRY (1998). *The Power of Discourse: An Introduction to Discourse Analysis*. Mahwah: Lawrence Erlbaum Associates.

CHUANG, JASON *ET AL*. (2015). "TopicCheck: Interactive Alignment for Assessing Topic Model Stability." *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 175–184.

COHEN, DANIEL & ROY ROSENZWEIG (2006). *Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web*. Philadelphia: University of Pennsylvania Press.

CRAIG, HUGH & ARTHUR F. KINNEY (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: University Press.

DAOUD, ADEL & SEBASTIAN KOHL (2016). *How Much Do Sociologists Write About Economic Topics?: Using Big Data to Test Some Conventional Views in Economic Sociology, 1890 to 2014* [MPIfG Discussion Paper 16/7]. Köln. Max-Planck-Institut für Gesellschaftsforschung. <http://www.mpifg.de/pu/mpifg_dp/dp16-7.pdf> [2019-12-20]

DAS, SANJIV R. & MIKE Y. CHEN (2007). "Yahoo! for Amazon. Sentiment Extraction from Small Talk on the Web." *Management Science* 53.9: 1375–1388.

DIMAGGIO, PAUL J., MANISH NAG, & DAVID M. BLEI (2013). "Exploiting Affinities Between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding." *Poetics* 41: 570–606.

EGGERS, ANDREW C. & ARTHUR SPIRLING (2014). "Ministerial Responsiveness in Westminster Systems: Institutional Choices and House of Commons Debate, 1832–1915." *American Journal of Political Science* 58.4: 873–887.

EISENSTEIN, JACOB *ET AL*. (2010). "A Latent Variable Model for Geographic Lexical Variation." *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. 1277–1287.

ERLIN, MATT (2014). "The Location of Literary History: Topic Modeling, Network Analysis, and the German Novel, 1731-1864." *Distant Reading: Topologies of German Culture in the Long Nineteenth Century*. Eds. Matt Erlin & Lynne Tatlock. Rochester/New York: Camden House. 55–81.

ERLIN, MATT & LYNNE TATLOCK (2014). "Introduction: "Distant Reading" and the Historiography of Nineteenth-Cenutry German Literature." *Distant Reading: Topologies of German Culture in the Long Nineteenth Century*. Eds. Matt Erlin & Lynne Tatlock. Rochester/New York: Camden House. 1–25.

FAIRCLOUGH, NORMAN (2007). Language and Power [Second Edition]. Harlow: Longman.

FAIRCLOUGH, NORMAN & RUTH WODAK (1997). "Critical Discourse Analysis." *Discourse as Social Interaction*. Ed. Teun van Dijk. London: SAGE Publications. 258–284.

FARRELL, JUSTIN (2016). "Corporate Funding and Ideological Polarization about Climate Change." *Proceedings of the National Academy of Sciences of the United States of America* 113.1: 92–97.

FEINERER, INGO, KURT HORNIK & DAVID MEYER (2008). "Text Mining Infrastructure in R." *Journal of Statistical Software* 25.5: 1–54.

GENOVESE, FEDERICA (2015). "Politics Ex Cathedra: Religious Authority and the Pope in Modern International Relations." *Research and Politics* 2.4: 1–15.

GOLDER, SCOTT A. & MICHAEL W. MACY (2011). "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures." *Science* 333: 1878–1881.

GOLDSTONE, ANDREW & TED UNDERWOOD (2014). "The Quiet Transformation of Literary Studies: What Thirteen Thousand Scholars Could Tell Us." *New Literary History* 45.3: 359–384.

GRAHAM, SHAWN, IAN MILLIGAN & SCOTT WEINGART (2016). *Exploring Big Historical Data: The Historian's Macroscope*. London: Imperial College Press.

GRIFFITHS, THOMAS L. & MARK STEYVERS (2004). "Finding Scientific Topics." *Proceedings of the National Academy of Science of the United States of America* 101.1: 5528–5235.

GRIFFITHS, THOMAS L., MARK STEYVERS & JOSHUA B. TENENBAUM (2007). "Topics in Semantic Representation." *Psychological Review* 114.2: 211–244.

GRIMMER, JUSTIN (2010). "A Bayesian Hierarchical Topic Model for Political Texts. Measuring Expressed Agendas in Senate Press Releases." *Political Analysis* 18.1: 1–35.

GRIMMER, JUSTIN & BRANDON STEWART (2013). "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21.3: 267–297.

HALL, DAVID, DANIEL JURAFSKY & CHRISTOPHER D. MANNING (2008). "Studying the History of Ideas Using Topic Models." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 363–371.

HASLINGER, PETER (2006). "Diskurs, Sprache, Zeit, Identitat: Plädoyer für eine Erweiterte Diskursgeschichte." *Historische Diskursanalysen: Genealogie, Theorie, Anwendungen*. Ed. Franz X. Eder. Wiesbaden: VS Verlag für Sozialwissenschaften. 27–50.

HOLMES, DAVID I. (1998). "The Evolution of Stylometry in Humanities Scholarship." *Literary and Linguistic Computing* 13.3: 111–117.

HORTON, RUSSEL *ET AL*. (2009). "Mining Eighteenth Century Ontologies. Machine Learning and Knowledge Classification in the Encyclopedie." *Digital Humanities Quarterly* 3.2. <http://www.digitalhumanities.org/dhq/vol/3/2/000044/000044.html> [2019-12-20]

HU, YUENING, JORDAN BOYD-GRABER, BRIANNA SATINOFF, & ALISON SMITH (2014). "Interactive Topic Modeling." *Machine Learning* 95.3: 423–469.

JACOBI, CARINA, WOUTER VAN ATTEVELDT & KASPER WELBERS (2016). "Quantitative Analysis of Large Amounts of Journalistic Texts Using Topic Modelling." *Digital Journalism* 4.1: 89–106.

JAMAL, AMANEY A. *ET AL*. (2015). "Anti-americanism and Anti-interventionism in Arabic Twitter Discourses." *Perspectives on Politics* 13.1: 55–73.

JELVEH, ZUBIN, BRUCE KOGUT & SURESH NAIDU (2014). "Detecting Latent Ideology in Expert Text: Evidence from Academic Papers in Economics." *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. 1804–1809.

LAFFERTY, JOHN D., & DAVID M. BLEI. (2006). "Correlated Topic Models." *Advances in Neural Information Processing Systems 18*. Eds. Yair Weiss, Bernhard Schölkopf & John C. Platt. MIT Press. 147–154.

JOHNSTONE, BARBARA (2002). *Discourse Analysis*. Malden: Blackwell Publishing.

JÄGER, SIEGFRIED (1999). "Einen Königsweg Gibt es Nicht: Bemerkungen zur Durchführung von Diskursanalysen." *Das Wuchern der Diskurse: Perspektiven der Diskursana-*

lyse Michel Foucaults. Eds. Hannelore Bublitz, *et al*. Frankfurt am Main/New York: Campus Verlag. 136–147.

JÄGER, SIEGFRIED (2012). *Kritische Diskursanalyse: Eine Einführung*. Münster: UNRAST-Verlag.

JÄGER, MARGARETE (1996). *Fatale Effekte. Die Kritik am Patriarchat im Einwanderungsdiskurs*. Duisburg: Duisburger Inst. f. Sprach- u. Sozialforschung.

JÄGER, SIEGFRIED & MARGARETE JÄGER (2007). *Deutungskämpfe: Theorie und Praxis Kritischer Diskursanalyse*. Wiesbaden: VS Verlag für Sozialwissenschaften.

JÄGER, SIEGFRIED & FLORENTINE MAIER (2009). "Theoretical and Methodological Apsects of Foucauldian Critical Discourse Analysis and Dispositive Analysis." *Methods of Critical Discourse Analysis*. Eds. Ruth Wodak & Michael Meyer. London: SAGE Publications. 34–61.

JÄGER, SIEGFRIED & JENS ZIMMERMANN (2010). *Lexikon Kritische Diskursanalyse: Eine Werkzeugkiste*. Münster: UNRAST-Verlag.

JOSHI, MAHESH *ET AL*. (2010). "Movie Reviews and Revenues: An Experiment in Text Regression." *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 293–296.

KELLER, REINER (2006). "Wissen oder Sprache?: Für Eine Wissensanalytische Profilierung der Diskursforschung." *Historische Diskursanalysen: Genealogie, Theorie, Anwendungen. Ed. Franz X. Eder*. Wiesbaden: VS Verlag für Sozialwissenschaften. 51–69.

KING, GARY (2009). "The Changing Evidence Base of Social Science Research." *The Future of Political Science: 100 Perspectives*. Eds. Gary King, Kay L. Scholzman, & Norman Nie. New York: Routledge. 91–93.

KING, GARY & WILL LOWE (2003). "An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders. A Rare Events Evaluation Design." *International Organization* 57.3: 617–642.

KING, GARY, JENNIFER PAN & MARGARET E. ROBERTS (2013). "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107.2: 1–19

KRESS, GUNTHER (1990). "Critical Discourse Analysis." *Annual Review of Applied Linguistics* 11: 84–99.

LANDWEHR, ACHIM (2001). *Geschichte des Sagbaren: Einführung in die Historische Diskursanalyse*. Tübingen: Edition Diskord.

LANDWEHR, ACHIM (2006). "Diskursgeschichte als Geschichte des Politischen." *Foucault: Diskursanalyse in der Politik*. Wiesbaden: VS Verlag für Sozialwissenschaften. Eds. Brigitte Kerchner & Silke Schneider. 104–122.

LAW, DAVID S. (2016). *Constitutional Archetypes (Legal Studies Research Paper Series No. 16-02-01*. St. Louis: Washington University in St. Louis, School of Law.

LAZER, DAVID *ET AL*. (2009). "Computational Social Science." *Science* 323: 721–723.

LOUGHRAN, TIM & BILL MCDONALD (2011). "When is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-ks." *Journal of Finance* 66.1: 35–65.

LUCAS, CHRISTOPHER *ET AL*. (2015). "Computer Assisted Text Analysis for Comparative Politics." *Political Analysis* 23.2: 1–24.

MARSHALL, EMILY A. (2013). "Defining Population Problems: Using Topic Models for Cross-National Comparison of Disciplinary Development." *Poetics* 41: 701–724.

MCAULIFFE, JON D. & DAVID M. BLEI. (2008). "Supervised Topic Models." *Advances in Neural Information Processing Systems 20*. Eds. John C. Platt, Daphne Koller, Yoram Singer, & Sam T. Roweis. Curran Associates. 121–128.

METAXAS, PANAGIOTIS T., ENI MUSTAFARAJ & DANIEL GAYO-AVELLO (2011). "How (Not) to Predict Elections." *Privacy, Security, Risk and Trust (PASSAT), IEEE Third International Conference on Social Computing (SocialCom)*. 165–171.

MILLER, IAN M. (2013). "Rebellion, Crime and Violence in Qing China, 1722-1911: A Topic Modeling Approach." *Poetics* 41: 626–649.

MIMNO, DAVID (2012A). "Computational Historiography: Data Mining in a Century of Classics Journals." *Journal on Computing and Cultural Heritage* 5.1: 1–19.

MIMNO, DAVID (2012B). *Topic Regression*. University of Massachusetts, Amherst. <http://scholarworks.umass.edu/open_access_dissertations/520/> [2019-12-20]

MIMNO, DAVID & ANDREW MCCALLUM (2008). "Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression." *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*. 411–418.

MIMNO, DAVID *ET AL.* (2011). "Optimizing Semantic Coherence in Topic Models." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 262–272.

MOGASHOA, TEBOGO (2014). "Understanding Critical Discourse Analysis in Qualitative Research." *International Journal of Humanities Social Sciences and Education* 1.7: 104–113.

MOHR, JOHN W. & PETKO BOGDANOV (2013). "Introduction - Topic models: What They Are and Why They Mattter." *Poetics* 41: 545–569.

MOHR, JOHN W. *ET AL.* (2013). "Graphing the Grammar of Motives in National Security Strategies: Cultural Interpretation Automated Text Analysis and the Drama of Global Politics." *Poetics* 41: 670–700.

MORETTI, FRANCO (2002). "Conjectures on World Literature." *New Left Review* 1:54–68.

NERBONNE, JOHN (2009). "Data-Driven Dialectology." *Language and Linguistics Compass* 3.1: 175–198.

NEWMAN, DAVID J. & SHARON BLOCK (2006). "Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper." *Journal of the American Society for Information Science and Technology* 57.6: 753–767.

O'CONNOR, BRENDAN *ET AL.* (2010). "A Mixture Model of Demographic Lexical Variation." *Proceedings of NIPS Workshop on Machine Learning for Social Computing*.

PARKER, IAN (1992). *Discourse Dynamics: Critical Analysis for Social and Individual Psychology*. London/New York: Routledge.

QUINN, KEVIN M. *ET AL.* (2010). "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54.1: 209–228.

REISIGL, MARTIN & RUTH WODAK (2009). "The Discourse-Historical Approach (DHA)." *Methods of Critical Discourse Analysis*. Eds. Ruth Wodak & Michael Meyer. London: SAGE Publications. 87–121.

RIDDELL, ALLEN B. (2014). "How to Read 22,198 Journal Articles: Studying the History of German Studies with Topic Models." *Distant Reading: Topologies of German Culture in the Long Nineteenth Century.* Eds. Matt Erlin & Lynne Tatlock. Rochester/New York: Camden House. 91–114.

ROBERTS, MARGARET E., BRANDON STEWART & EDOARDO M. AIROLDI (2016). "A model of text for experimentation in the social sciences." *Journal of the American Statistical Association*. 111.515: 988-1003. <http://scholar.princeton.edu/bstewart/publications/model-text-experimentation-social-sciences> [2019-12-20]

ROBERTS, MARGARET E. & BRANDON STEWART & DUSTIN TINGLEY (2014). "stm: R Package for Structural Topic Models," *Working Paper 176291*, Harvard University OpenScholar.

ROBERTS, MARGARET E., BRANDON STEWART & DUSTIN TINGLEY (2016). "Navigating the Local Modes of Big Data: The Case of Topic Models." *Computational Social Science: Discovery and Prediction*. Ed. Michael Alvarez. Cambrigde: Cambridge University Press. <http://scholar.harvard.edu/files/dtingley/files/multimod.pdf> [2019-12-20]

ROBERTS, MARGARET E., BRANDON STEWART, DUSTIN TINGLEY & EDOARDO M. AIROLDI (2013). "The Structural Topic Model and Applied Social Science." *Advances in Neural Information Processing Society*. Prepared for the NIPS 2013 Workshop on Topic Models: Computation, Application, and Evaluation <http://scholar.princeton.edu/files/bstewart/files/stmnips2013.pdf> [2019-12-20]

ROBERTS, MARGARET E. *ET AL.* (2014). "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58.4: 1064–1082.

ROSEN-ZVI, MICHAEL *ET AL*. (2010). "Learning Author-Topic Models from Text Corpora." *ACM Transactions on Information Systems* 28.1: 1–38.

SARASIN, PHILIPP (2003). *Geschichtswissenschaft und Diskursanalyse*. Frankfurt am Main: Suhrkamp.

SCHÄFFNER, CHRISTINA (1996). "Editorial." *Discourse and Ideologies*. Eds. Christina Schäffner & Helen Kelly-Holmes. Clevedon: Multilingual Matters. 1–6.

SCHMIDT, BENJAMIN M. (2012). "Words Alone: Dismantling Topic Models in the Humanities." *Journal of Digital Humanities* 2.1. <http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/> [2019-12-20]

SHARMA, SUGAM *ET AL*. (2014). "A Brief Review on Leading Big Data Models." *Data Science Journal* 13.4: 138–157.

SHELLMAN, STEPHEN M. (2008). "Coding Disaggregated Intrastate Conflict. Machine Processing the Behavior of Substate Actors over Time and Space." *Political Analysis* 16.4: 464–477.

SLAPIN, JONATHAN B. & SVEN-OLIVER PROKSCH (2009). "How to Avoid Pitfalls in Statistical Analysis of Political Texts: The Case of Germany." *German Politics* 18.3: 323–344.

SPANGHER, ALEXANDER (2015, August 11). "Building the Next New York Times Recommendation Engine." *The New York Times*. <http://open.blogs.nytimes.com/2015/08/11/building-the-next-new-york-times-recommendation-engine/> [2019-12-20]

STEDE, MANFRED (2007). *Korpusgestützte Textanalyse: Grundzüge der Ebenen-orientierten Textlinguistik*. Tübingen: Narr Francke Attempto.

STEPHENS-DAVIDOWITZ, SETH I. (2012). "The Effects of Racial Animus on a Black Presidential Candidate. Using Google Search Data to Find What Surveys Miss." *SSRN*. <http://dx.doi.org/10.2139/ssrn.2050673> [2019-12-20]

TANGHERLINI, TIMOTHY R. & PETER LEONARD (2013). "Trawling in the Sea of the Great Unread: Sub-corpus Topic Modeling and Humanities Research." *Poetics* 41: 725–749.

TAUSCZIK, YLA R. & JAMES W. PENNEBAKER (2009). "The Psychological Meaning of Words. LIWC and Computerized Text Analysis Methods." *Journal of Language and Social Psychology* 29.1: 24–54.

TETLOCK, PAUL C. (2007). "Giving Content to Investor Sentiment: The Role of Media in the Stock Market." *The Journal of Finance* 62.3: 1139–1168.

VAN DIJK, TEUN (1997). "The Study of Discourse." *Discourse as Structure and Process 1*. Ed. Teun van Dijk. London: SAGE Publication. 1–34.

VAN DIJK, TEUN (2009). "Critical Discourse Studies: A Sociocognitive Approach." *Methods of Critical Discourse Analysis*. Eds. Ruth Wodak & Michael Meyer. London: SAGE Publications. 62–86.

WALLACH, HANNA M., DAVID MIMNO & ANDREW MCCALLUM (2009). "Rethinking LDA: Why Priors Matter." *Advances in Neural Information Processing Systems* 22: 1973–1981.

WANG, XUERUI & ANDREW MCCALLUM (2006). "Topics over Time: A Non-markov Continuous-Time Model of Topical Trends." *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 424–433.

WEBBER, BONNIE & ARAVIND JOSH (2012). "Discourse Structure and Computation: Past, Present and Future." *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*. 42–54.

WITTEK, PETER & WALTER RAVENEK (2011). "Supporting the Exploration of a Corpus of 17th-Century Scholarly Correspondences by Topic Modeling." *Supporting Digital Humanities 2011: Answering the Unaskable. Ed. Bente Maegaard. Kopenhagen*. <http://www.clarin.nl/sites/default/files/sdh2011-wittek-ravenek.pdf> [2019-12-20]

WODAK, RUTH & CHRISTOPH LUDWIG (1999). "Introduction." *Challenges in a Changing World: Issues in Critical Discourse Analysis*. Eds. Ruth Wodak & Christoph Ludwig. Wien: Passagen-Verlag. 11–19.

WODAK, RUTH & MICHAEL MEYER (2009). "Critical Discourse Analysis: History, Agenda, Theory and Methodology." *Methods of Critical Discourse Analysis.* Eds. Ruth Wodak & Michael Meyer. London: SAGE Publications. 1–33.