Topic models do not model topics: epistemological remarks and steps towards best practices

Anna Shadrova

Journal of Data Mining and Digital Humanities

Topic models do not model topics: epistemological remarks and steps towards best practices

Anna Shadrova[1]

[1]Humboldt-Universität zu Berlin, Germany

Corresponding author: Anna Shadrova , anna.shadrova@hu-berlin.de

Abstract

The social sciences and digital humanities have recently adopted the machine learning technique of topic modeling to address research questions in their fields. This is problematic in a number of ways, some of which have not received much attention in the debate yet. This paper adds epistemological concerns centering around the interface between topic modeling and linguistic concepts and the argumentative embedding of evidence obtained through topic modeling. It concludes that topic modeling in its present state of methodological integration does not meet the requirements of an independent research method. It operates from relevantly unrealistic assumptions, is non-deterministic, cannot effectively be validated against a reasonable number of competing models, does not lock into a well-defined linguistic interface, and does not scholarly model topics in the sense of *themes* or *content*. These features are intrinsic and make the interpretation of its results prone to apophenia (the human tendency to perceive random sets of elements as meaningful patterns) and confirmation bias (the human tendency to perceptually prefer patterns that are in alignment with pre-existing biases). While partial validation of the *statistical* model is possible, a conceptual validation would require an extended triangulation with other methods and human ratings, and clarification of whether statistical distinctivity of lexical co-occurrence correlates with conceptual topics in any reliable way.[1]

Keywords

Topic modeling; digital humanities; information extraction for scientific inquiry

I INTRODUCTION

Topic modeling is a machine learning technique that classifies words into so-called *topics* and computes an estimated proportion of those for documents and corpora (collections of text). More recent implementations also allow for the analysis of topic correlations with text metadata, such as changes over time or distributions across authors (Blei and Lafferty, 2006; Roberts et al., 2019, 2014). It was initially developed and continues to be used for information retrieval and text classification purposes in applied contexts (Bao et al., 2009; Asuncion et al., 2010; Ramage et al., 2011; Wang and Blei, 2011; Chuang et al., 2013; Si et al., 2014; Zhong et al., 2015; van Der Hooft et al., 2016; Liu et al., 2016; Boyd-Graber et al., 2017; Kuhn, 2018; Liu et al., 2019; Reber, 2019; Korfiatis et al., 2019, and many others). More recently, it has also gained momentum in the context of so-called *distant reading*[2] in the digital humanities and

[1]I am grateful to Frédéric Clavert and an anonymous reviewer for their very constructive and helpful comments. All remaining flaws and errors are of course my own.

[2]*Distant reading*, as opposed to *close reading*, describes the analysis of surface patterns retrieved from large corpora from leveraging computational techniques of text mining. For applications and critical contributions to the

social sciences, where it is now increasingly being used to answer subject-specific research questions regarding the distributions of content in literary text (Asgari et al., 2013; Tangherlini and Leonard, 2013; Jockers and Mimno, 2013; Underwood, 2014; Goldstone and Underwood, 2014; Weitin and Herget, 2017; Mitrofanova and Sedova, 2017; Schöch, 2017; Erlin, 2018; Navarro-Colorado, 2018; Jacobs, 2018; Sieg, 2019; Dahllöf and Berglund, 2019; Liu and Jin, 2020), court decisions (Livermore et al., 2016; Panagis et al., 2016; Carter et al., 2016; Law, 2016; Wang et al., 2017; Rice, 2017; Young, 2019; Lampach and Dyevre, 2018), political and legal debate (Young, 2012; Greene and Cross, 2016; Sterling et al., 2019; Grimmer, 2010), media coverage (DiMaggio et al., 2013; Bertalan and Ruiz, 2019; Chandelier et al., 2018; Yang et al., 2011), or academic journals (Chen et al., 2020; Lindstedt, 2019; Wang et al., 2017).[3]

This extension is epistemologically problematic for a number of reasons. While issues around the subjectivity of topic labeling and a lack of mathematical validation of topic models have been discussed since the beginning of their application (Chang et al., 2009; Ramirez et al., 2012; Arora et al., 2013; Chuang et al., 2013, 2015), some deeper problems remain largely unaddressed so far, although there has been some work showing the practical issues (overly mixed, uninterpretable, incomplete topics etc.) as they occur if the data is not sufficiently well prepared (Schmidt, 2012; Boyd-Graber et al., 2014). The argument in this paper concerns structural and conceptual properties of topic modeling and its linguistic underpinnings, specifically

a) the underlying model of a latent, unshifting and uniquely definable semantic space at the heart of topic modeling that is greatly at odds with linguistic and text research theory. Topics, themes, or categories of content are not well-defined linguistic concepts, their boundaries are generally fuzzy and dynamic, and they can only be *con*structed in an in-depth analytical process, not *ex*tracted as basic information, because they are largely derived from implicit meaning and not word co-occurrence *per se*;
b) the fact that statistical word distributions, may they be more or less distinct, do not guarantee topic distinctivity from a linguistic perspective, which means that even statistically well-defined and reliable models cannot ensure concept validity;
c) the heuristic, non-deterministic character of topic modeling, that is at odds with the perceived objectivity of its output, while the large number of models for any given corpus effectively escapes validation between a representative proportion of possible models;
d) the estimation of topic prevalence in documents through statistical means, that is an oversimplification of the complexities of the quantification of meaning and is further complicated by text-linguistic features such as text length, complexity, or genre;
e) the clash between the actually narrow evidential scope of topic modeling results, i.e., the fact that they would require much methodological effort and contextualization to be integrated into a scholarly argument in somewhat conclusive ways, with a common practice to include results as "exploratory" without further integration.

Points a) - d) are equally true of topic modeling in applied contexts, for which it was initially developed. However, in those fields, results are subject to immediate external validation, for example through changed customer behavior in a recommender system, or the actual synthesis of hypothetically possible molecules, cf. van Der Hooft et al. (2016). Unlike this, themes and categories of content *themselves* are part of the final output of text-based research in academic contexts. They cannot usually be externally verified or validated, putting the full weight of their

debate, see Moretti (2013); Ascari (2014); Jänicke et al. (2015); Underwood (2017); Drucker (2017).

[3]More references and a mapping of technologies to research questions can be found in McFarland et al. (2013). Boyd-Graber et al. (2017) put together a 150-page-long and rich overview of applications and their technical context.

acceptability on their argumentative embedding into theoretical and/or empirical models upon which future research can build. *Topic modeling does not provide models of this kind.*

The purpose of this paper is to provide a systematic review of the issues mentioned from the research perspectives of quantitative corpus linguistics, qualitative text-based research fields, and epistemology, in order to demarcate scopes of application as well as epistemological limitations.

It begins with a brief discussion of what constitutes a *topic, theme,* or *content* in different subdisciplines of text-based research fields such as linguistics, literary studies, and social sciences. It then presents arguments for why topic modeling is not an automation of the process of category construction by first providing a conceptual and largely non-technical introduction into topic modeling (section III); and then a more thorough discussion of the type of evidence derived from topic modeling and how it a) clashes with linguistic concepts and b) is difficult to integrate into scholarly arguments (section IV).

This is relevant in the context of the digital humanities and text-based research fields shifting from more qualitative to more quantitative work in recent years, where methodological sophistication has not kept up with the evolution of technology. Since approaches towards more structured evaluation, systematization, and validation of methods are necessary for further development, some suggestions to this effect are made in section V.

II WHAT ARE TOPICS AND HOW CAN THEY BE MODELED?

As competent adult speakers of a language, we often have a good idea of what a text is about – even if it is a very short text, like a single sentence, or one that is syntactically deprecated, like a newspaper headline ("Trump chief of staff Meadows under fire for handling of pandemic and other crises", The Washington Post online, 2020/10/27[4]). We derive this idea, i.e., the subject, topic, theme, aboutness, or content of a text, partly from the words it contains, and partly from our taxonomic and ontological understanding of the world. For example, in the headline above, we can infer certain aspects of the meaning from words such as *chief of staff, pandemic,* or *crises*. But in order to correctly extrapolate the topic, we also need the context of current world politics (for example, who is *Trump*), and an understanding of the contextualized meaning of *under fire* (in a professional setting as opposed to a military siege) or *handling* (a pandemic as opposed to a zoo animal). However, even with this knowledge, defining an appropriate topic label is still not trivial. Depending on the underlying ontology and the goal of the classification, it could range from *US politics* to *a specific period in the life of Mark Meadows* to *world events causing professional crises*.

There are at least four largely separate branches of scholarly research that work with the idea of topic, theme, or category of content: linguistics, literary studies, information science, and the social sciences. Their respective ideas of what constitutes a topic are pragmatically and theoretically diverse – in linguistics, the notion of a *topic* is relevant largely in the context of information structure and denotes unit of information in a sentence or discourse that a proposition pertains to, or a referent that something relevant is said about; in literary studies, a *theme* is an implicit story or larger category of meaning derived from the words of a work, and is webbed into a historical or societal context; in the social sciences, a *topic* or *category of content* is implicit or explicit meaning derived from words that belong to similar semantic categories. Information science further uses the terms *subject* or *aboutness*, describing overarching labels of classification that facilitate effective search in libraries or search engines.

[4] https://www.washingtonpost.com/politics/meadows-trump-coronavirus/2020/10/26/475f03d2-122f-11eb-82af-864652063d61_story.html

For example, in the sentence *The King died and then the Queen died*, a linguist might say that the sentence topic is *King* in the first clause and *Queen* in the second; a literary scholar might say the theme is *loss* or *grief*; and a social scientist might say the topic is *death*, *royalty*, or *the history of a certain country at a certain time*. The set of potential categories depends on the contextualization by the scholar within the (potential) taxonomy of each subject.

However, the scope of such a contextualization or categorization is not uniquely defined, and the concept of a topic, theme, etc. itself is further debated across disciplines. For linguistics, Goutsos (1997, 1) notes that

> "[i]n reviewing the work on topic and theme, one is struck by the almost total lack of consensus among linguists regarding the nature, the defining characteristics, and the scope of application of the notions employed. The only opinion that seems to be widely shared is that the area (...) is riddled with problems. One of the reasons for this must be the continuously proliferating literature. It is a difficult task to review even the existing reviews (...). This persisting interest is certainly indicative of the importance of the area. Yet it is quite doubtful whether the terms as used have left the realm of intuition and acquired a precise meaning".

The author continues to name at least 15 definitions covering a range from "a recognizable unit (of ideas, words, etc.)" to "(...) a unifying thread running through the text as a whole". Schlobinski and Schütze-Coburn (1992, 114) and Kehler (2004, 238) even suggest that topic is an epiphenomenon arising from textual coherence with no linguistic reality of its own. A similar debate with strikingly similar results has occurred in thematics, a subdiscipline of literary studies – what makes a theme, a motif, the subject, and so on, is not uniquely defined in the field but rather attracts diverging and often contradictory concepts (Van Peer (2002), Pettersson (2002, 238), Rimmon-Kenan (1995)) or is even questioned with respect to the value of its contribution in the first place (Sollors, 2002).

This abundance of approaches stems from the fact that topics (themes, subjects, etc.) are not a linguistic category per se – much of what topic or content analysis is about is not included in actual words or their co-occurrence. As Krippendorff (2004a, 10) reflects for content analysis in the social sciences:

> "Content analysts must predict or infer phenomena that they cannot observe directly. The inability to observe phenomena of interest tends to be the primary motivation for using content analysis. Whether the analyzed source has reasons to hide what the analyst desires to know (...) or the phenomena of interest are inaccessible in principle (e.g., an individual's attitudes or state of mind, or historical events) or just plain difficult to assess otherwise (such as what certain mass-media audiences could learn from watching TV), the analyst seeks answers to questions that go outside a text."

Topics, in most definitions, are hence not *direct derivatives of the data* but *interpretations of aspects of the data in context*, where the context is provided by the scholar's knowledge of a subject and taxonomies and debates in the field. Topic or content analysis in the scholarly sense is then not a process of *ex*traction of existing entities, but one of category *con*struction.[5] The question of whether topics are *objectively* included in any data can hence only be answered in

[5]Even if data is mapped to existing categories in the field, this still requires the construction of the same categories from the data.

a process-oriented way. If a certain type of function, i.e., method, takes a certain input and yields a certain output, then relative to the function, it can be said that the output is justified. Since the function is crucial for the validity of results, each discipline has created an extensive methodology of qualitative research for questions of this kind.[6] Approaches like hermeneutics or grounded theory may not usually be considered as highly precise in the same way that quantitative approaches are, but they do place strong emphasis on the role of contextualization through existing taxonomies, influence of researcher subjectivity, and cognitive biases such as confirmation bias (the tendency of the human mind to perceptually prefer patterns that are in alignment with existing expectations) or apophenia (reading patterns into random data, like cloud animals, cf. Dixon (2012)). The rest of the paper aims to show how topic modeling is *not* an automation of such kinds of analysis, and that the same caveats persist and are in fact amplified through a massive and consequential reduction of information in topic modeling.

III WHAT IS TOPIC MODELING?

Topic modeling is an unsupervised machine learning technique that classifies words across documents into co-occurring word bundles, so-called *topics*. *Unsupervised* means that the classification is based only on the corpus content itself; no further information, such as a gold standard of labeled content, is provided. There is, however, usually some preprocessing involved in order to avoid trivial or random topics. This includes the filtering of highly frequent words such as articles or prepositions, and setting all words to lowercase (this can have repercussions on the content in languages like German, where nouns are capitalized). Most implementations offer the option to specify other words that should not be considered in the model, for instance words that might skew results due to ambiguity or overspecificity.

The statistical model extracted from a corpus of texts builds on the assumption that there exists a limited and fixed number of recurrent topics, i.e., "things that can be talked about" in the world,[7] and that this is reflected in corpora through the frequent co-occurence of topic-specific words in separate texts. For example, some words that might frequently co-occur are *health, doctor,* and *hospital*, suggesting a topic of *medicine* or *health care*. Each document is modeled as a stochastically distributed mixture of only a few topics. For example, it is more likely that a *medicine* topic is bundled with a *human rights* topic than a *board games* topic, although both are technically possible; and no document typically contains all topics that exist in the model (rendering the probability of some topics near zero in some documents). More formally, the underlying model is a stochastic mixed-membership model of latent semantic spaces.

Topics are presumed to contain words at a certain and unchanging probability (words that are not relevant to a topic have a probability of nearly zero). Thus, *topics* in topic modeling are distributions of words, and are probabilistically distributed across documents. A *topic model* then is a distribution of distributions (of words).

These distributions can be computed with a number of algorithms: Latent Semantic Indexing (LSI) or Latent Semantic Analysis (LSA) (Deerwester et al., 1990), Non-Negative Matrix

[6] Content analysis: Krippendorff (2004a); Macnamara et al. (2005); Kimberly (2011); Mayring (2004); Graneheim and Lundman (2004); Krippendorff (2004b); Thematics in literary studies: Hasan (1967); Rimmon-Kenan (1995); Pettersson (2002).

[7] "Common to all unsupervised topic models is the idea that language is organized by latent dimensions that actors may not even be aware of. When applied to everyday speech, basic (unsupervised) topic models usually identify areas of discussion – like driving and stop signs, and distinguish that from, say, dating", McFarland et al. (2013, 5). It is interesting that the author refers to everyday subjects that are relatively clearly defined and distinguishable, much unlike the fuzzy and gradient categories of scholarly research.

Factorization (NMF) (Kuang et al., 2015), and, most widely used at present, Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and modifications thereof.[8] Without going into detail, a rough description is in order to prepare the following line of argument.[9]

Two distinct sets of algorithms are used in topic modeling. One is based on the dimensionality reduction of vector spaces (Latent Semantic Indexing/Analysis (LSI/LSA), Non-Negative Matrix Factorization (NMF)). These are deterministic, which means that they will always yield the same model for the same data, no matter how often the algorithm is run or which word it uses as its starting point. Those algorithms use different types of dimensionality reduction, but both identify topics based on a word co-occurrence matrix (also called a document-term-matrix). That is a matrix containing the frequency of co-occurrence of each word with all other words for each document, which is then decomposed into a vector of words pertaining to each topic, and another vector of topics pertaining to each document. From these vectors, similarity is computed, typically through cosine distance. Think of vectors as arrows in a two-dimensional space. If two arrows point in different directions, their roots will meet at an angle that can be used for comparison: the more similar the direction, the smaller the angle, with a maximum difference of 180°. Vectors in topic modeling are not two-dimensional, but the same principle applies.

With these algorithms, the number of topics is derived from the matrix, and the same matrix always yields the same topics in both composition and distribution. LSI/LSA and NMF were the first kinds of topic models, but have largely been superseded by the more advanced algorithm Latent Dirichlet Allocation (LDA). In LDA, the number of topics is not derived from a document-term-matrix, but chosen manually. Based on the desired number of topics, LDA first assigns words to topics at random and then, aiming to maximize a distinctivity function, iteratively swaps words between topics. The final model is one that, depending on the initialization point, yields *the most distinct or separable distribution* of topics.[10]

This is in line with the underlying (albeit problematic) assumption that topics are discrete and distinct, which is to say there is no continuity or gradient between topics; that they are marked by specific words; and that topics differ by document – while some very frequent topics may occur in most documents at varying rates, overall, topics are supposed to work as a classifier and not be overly equally distributed.

Since the function is maximized to k chosen clusters iteratively starting from a point of initialization, LDA is not deterministic. Depending on the point of initialization (the initial random assignment of words to topics), results may and do vary. The function can be solved to any number of topics, in the same way that it is possible to divide the things one owns into any

[8] Probabilistic versions of LSA/LSI and NMF also exist. It appears, however, that those are mathematically equivalent and only represent special cases of LDA, see de Paulo Faleiros and de Andrade Lopes (2016); Girolami and Kabán (2003); Ding et al. (2006).

[9] A technical, but very approachable comparison of the main algorithms can be found here: https://medium.com/@souravboss.bose/comprehensive-topic-modelling-with-nmf-lsa-plsa-lda-lda2vec-part-1-20002a8e03ae. For a comparison of LDA and NMF, see Chen et al. (2019). For an overview of LDA and its applications in some technical detail, see Blei et al. (2010) and Blei (2012). For some modifications of LDA, see Yu et al. (2017).

[10] The name *Dirichlet* in LDA refers to the Dirichlet distribution named after the 19th century mathematician Johan Peter Gustav Lejeune Dirichlet. A Dirichlet distribution is a multivariate probability density distribution used in Bayesian statistics. A maximally clustered distribution, i.e., one where high probability is assigned to few members in a small range, is approximated in LDA through Bayesian mixed-membership modeling. The details are irrelevant for the argument in this paper and will not be further discussed.

number of (sufficiently large) drawers. Obviously, the quality or usefulness of the partitioning may suffer from too many or too few drawers, and the same goes for topics in LDA: there is no way of determining the ideal number of topics a-priori. Attempts to approximate the ideal number include the combination with other statistical approaches, such as a Principle Component Analysis (PCA), that determines the statistically best number of clusters, which can then be used as *k* for the number of topics; or through the addition of a Non-Negative Matrix Factorization into preprocessing (the latter is implemented in Arora et al. (2014)). However, neither Principle Component Analysis nor NMF definitively yields the best separation of topics – this is obvious from the fact that it has been deprecated in favor of LDA due to frequently unsatisfactory results. LDA, overall, yields much more impressive results compared to older algorithms in terms of the interpretability and usability of topics in information retrieval contexts – but it cannot detect the overall 'best' model without further input.

While more can be said about the technical and mathematical aspects of various topic modeling algorithms, for the sake of the argument in this paper it suffices to know that

- *topics* are distributions of words,
- *topic models* are distributions of distributions (of words) across documents,
- topic models are *mixed-membership models*, meaning that several topics can be represented in a document and words can occur in several topics,
- the most frequently used topic modeling algorithm, LDA, is *probabilistic*, and *maximizes the distinctivity* of topics, i.e., distributes words across topics in such a way that topics are most clearly distinct from one another,
- the *number of topics* in LDA *is chosen manually*,
- *LDA is heuristic*, not deterministic: The same data can and does yield different models for the same number of topics if a different point of initialization is chosen.

IV WHAT KIND OF EVIDENCE DO TOPIC MODELS PROVIDE?

There are several intrinsic aspects of topic modeling in its current implementations that limit the scope of the evidence it presents:

1. Topic modeling is incomplete, heuristic, and escapes validation, it does hence not provide a unique or the objectively best model of corpus content even statistically;
2. Topic modeling operates from a massive and unpredictable linguistic dimensionality reduction and relevantly unrealistic assumptions, it does hence not provide conceptual validity with respect to the linguistic reality;
3. Topic modeling does not reliably quantify meaning, its quantitative output (estimation of topic prevalence in documents and by metadata, such as year or author) is hence not an exact or nearly exact measurement of content distribution;
4. Topic modeling does not allow for conscious qualitative parameter setting, it is hence rather limited in scholarly application;
5. Topics derived from topic modeling are still constructions, not objective observations.

These will be reviewed from a methodological and a linguistic perspective in some detail in the following sections.

4.1 Topic modeling is incomplete

Topic modeling can only reasonably include words of a certain frequency spectrum into the analysis. Including all lexemes in a topic model is generally ineffective due to high computational cost. It is also limited in terms of statistical relevance – an infrequent word occurring any number out of its limited total occurrences in the vicinity of more frequent words will not tip

the overall statistics compared to the higher clustering power of those frequent words.[11] While this may sound like it affects only a few words at the end of the frequency range, the opposite is in fact true. Only a very small proportion of lexemes occurs sufficiently frequently to make any reasonable statistical statement. In fact, words are distributed in a way that only few occur at high frequency (like English *the, to,* or *is*). Up to half of all words in large corpora occur only once (so-called *hapax legomena*, or simply hapaxes), another large section twice, three times, and so on. This is frequently modeled as a Zipf- or power law distribution (Baayen, 2002; Zipf, 1965) and entails that the number of *different* words (lexemes) used in corpora is very high. For computational efficiency, the R stm package (Roberts et al., 2019) works with the 10 000 most frequent lexemes. While this may cover a wide lexical range, it is far from complete: For example, in the corpus of German Federal Constitutional Court decisions containing 3312 decisions from the years 1951-2017 (Möllers et al., 2021), this leaves 48 000 lexemes, or 82.8%, unused.

This limitation also implies that topic modeling is *massively* (in fact, exponentially) more incomplete with respect to the included lexemes in larger corpora – more frequent lexemes reach saturation in relatively small corpora, while the number of hapaxes continues to grow. While hapaxes cannot be used for statistical analysis, the methodological implication that over 80% of the lexical material making up a corpus cannot be used for the estimation of its content distribution is certainly uncomfortable from a scholarly perspective. Even the exclusion of a single word from a topic can change its interpretation – consider for example the set of {*table, bar, box*} (perhaps labeled as *furniture*) vs. the same set with an additional *graph*, which may change it to *quantitative research results*. If most words are excluded, then the diversity of the resulting topics is necessarily limited.

It is further unknown that topic-relevant words are also distributed in a power law function, i.e., that a topic is made up from some prototypical and many other related, but infrequent words. It is possible that the most distinguishing words are all located within the lower frequency spectrum of the distribution. In fact, there is a lack of *any* quantitative model of within-topic distributions, rendering an estimation of the effect of incompleteness nearly impossible.

4.2 Topic modeling is heuristic and escapes validation

LDA is a heuristic that initializes topic distributions from one word and maximizes distinctivity from there. In other words, it takes a lexeme and distributes other lexemes according to the one it chose. This obviously depends on the initialization point: if we take any object in our home and put it in a box, thereby defining the box as a container for similar objects, and then arrange all other objects by co-occurrence, results will vary depending on the first object. If the first object was a hammock, we may find {*hammock, straw hat, lemonade*}, if we start from a sun chair, we may find {*sun chair, straw hat, lemonade*}, but not *hammock*. Both are distinct categorizations driven by the order of input and the classes that exist until that point, and neither is wrong. But which one is better? That depends on the goal of the classification, which topic modeling is blind to.

[11] Let it be noted that all statistical computations over words in a corpus are epistemologically problematic. The lexicon does likely not meet assumptions of ergodicity, i.e., path-independence (Debowski, 2018), and stationarity (Piantadosi, 2014), i.e., unchangeability over time and space, which are central to stochastic theory and statistics. In fact, it appears that the concept of a fixed probability of words is strange in too many ways to be considered epistemologically safely employable (Shadrova, in press). However, since topic modeling uses relative frequency largely descriptively, without recurrence to an external totality or expected values in the final result, those concerns will not be further addressed in this paper.

The above example is not entirely fair, because LDA would in fact likely sort the hammock into the same group via its co-occurrence with other items. It is possible for words to co-occur in the same topic even if not all of them occur in a document, i.e., for topics to contain mutually exclusive words. However, the overall computation relies on the maximization of distinctivity, which means that it will try to find the most distinct distribution as seen from a specific starting point, which necessarily results in different models for the same data. In addition, models are highly susceptible to changes in the data, in the same way that the distribution of things from our home in boxes will change if there is just one item that does not fit with any of the boxes we have defined. Empirically, this shows up as high sensitivity to minor changes in the corpus. For example, Wendel et al. (in press) find massive changes in topic distributions and prevalence between topic models of the corpus of German Federal Constitutional Court decisions based on over 3000 texts and the same corpus with another 10 texts added.

Since topic models vary depending on the order of the words they are fed, there are many different models, i.e., distributions of topics, for each corpus, and even more different topics. This is widely recognized as a problem in the application of topic modeling and usually tackled in one of three ways: (1) through fixed initialization points based on pre-processing, or through validation of the model by either (2) computational measures or (3) human rating (Boyd-Graber et al., 2014). All three are epistemologically problematic. Defining the initialization point based on pre-processing, for example by adding a Principal Component Analysis (PCA) or Non-Negative Matrix Factorization (NMF, as implemented in the R stm package (Roberts et al., 2019) with Arora et al. (2014)'s algorithm), estimates a good starting point and a statistically optimal number of topics. However, this is done based on algorithms that *by themselves* do not yield as convincing results (hence the development of LDA in the first place). While in the combination of LDA and NMF or PCA, topic coherence reaches more satisfactory levels in applied contexts than any of the individual procedures, the dimensionality reduction to n optimal topics cannot provide a conceptual guarantee for ideal topic distributions.

Solutions (2) and (3), human and machine ratings (metrics) of topic coherence (Chang et al., 2009; Lau et al., 2014; Bhatia et al., 2018; Wesslen, 2018; McFarland et al., 2013), are equally limited, since they can only be computed for an existing model. They can only offer an evaluation of whether a model reaches a minimum of coherence, not a validation of the model against other models. This marks a lower threshold, but that does not suffice for scholarly aims – it is generally not the aim of research to find any minimally coherent model, but one that is most descriptive of the data in light of a specific research question.

To fully validate the model for scholarly purposes, it would be necessary to compare all, or at least a representative proportion, of possible topic models of a corpus. Since there is a range of plausible topic numbers for any corpus, and unlike for information retrieval, the exact number of categories matters for taxonomic ordering, each of those topic numbers would require its own computation from all initialization points.

Let us say we want to compare models of 100–200 topics for a large corpus, incrementing in steps of 25 (100, 125, 150, 175, 200), that yields five models to compare – this is still rather coarse, the model that fits the data best might well be one of 127 or 138 topics. Using stm's 10 000 most frequent lexemes, a distinct topic model could be computed for each of those. For five topic models and 10 000 lexemes, there is a set of 50 000 potentially different topic models. Some of those may overlap, but it is difficult to predict which ones those will be or the scale or significance of their overlap.

Even if one were to invest the computational power and then were dedicated enough to manually check and cross-compare 50 of those models (1225 combinations of between 100 and 200 topics), that would *still leave an uncertainty of 99.9%*. Human validation then definitively becomes a lower threshold for acceptability of a more or less random model, not quality assurance in the sense of a choice of the best, or even one of the better, empirical models.

Consequently, it is impossible to validate topic models against their combinatorial power in practice, and we are factually forced to choose one from a random and small set that happens to be computed first. The only validation that can be provided for the full set of models that latently exist in the data is validation through metrics such as load, entropy, or other measures of coherence (McFarland et al., 2013; Boyd-Graber et al., 2014; Wesslen, 2018) – and even that is only an option if one has the computational power ready to compute thousands and thousands of topic models, which in practice is nearly impossible within regular research contexts. However, even then, statistical distinctivity of topics has no linguistic correlate: there is no concept in linguistics that would relate certain degrees of statistical distinctivity to certain qualitative aspects like goodness or coherence of topics. Neither preprocessing through PCA or NMF nor coherence metrics can thus provide concept validity.

4.3 Topic modeling is unlinguistic

Topic modeling works with word frequencies, but it is not a linguistic model. It is built on the idea that meaning is a latent structure formed from or expressed through co-occurring words, and that this meaning can occur in more or less distinct units that are statistically extractable. While there is certain common sense to the idea that meaning emerges from the combination of words, this alone does not suffice for an accurate and comprehensive linguistic description.

In topic modeling, the concept of *words working together to construct a topic* is represented by spatial proximity (typically within a document, but smaller sections are also possible), and frequently co-occurring words are presumed to form tighter or more coherent semantic groups than less frequently co-occurring words. Within linguistics, this idea is modeled in the subfield of distributional semantics (see Baroni et al. (2014) and Fabre and Lenci (2015) for an overview). It is widely used in computational linguistics, especially in applied contexts, for example in word embeddings (Peters et al., 2018; Devlin et al., 2018; Levy and Goldberg, 2014; Liu et al., 2015; Ethayarajh, 2019) or sentiment analysis (Medhat et al., 2014; Bakshi et al., 2016). However, the limitations of deriving higher-level information from word co-occurrences alone are also frequently discussed and lead to the prolific development of combinations with other models of meaning, such as visual information, knowledge graphs, and relational semantics (Herbelot, 2013; Bruni et al., 2014; Fried and Duh, 2014; Speer and Lowry-Duda, 2017; Lengerich et al., 2017; Thoma et al., 2017). This is due to the fact that what seems like a straight-forward model – *words that occur together construct larger units of coherent meaning* – is in fact a massive, consequential, and largely undpredictable linguistic dimensionality reduction. To illustrate only some of the problems as they occur in topic modeling:

1. **Statistical distinctivity does not equal thematic distinctivity, coherence, or granularity.** Topic modeling through LDA is based on a maximization function, suggesting that higher statistical distinctivity makes for better topics. *There is no concept in linguistics that would cover this*. It implies that distinctively co-occurring words *do* form semantic units (i.e., topics), while linguistics only states that co-occurring words *can* form semantic units, but does not specify this quantitatively in any way.
   In language, similar words do not necessarily denote similar topics, while different words in fact can. Rare words co-occurring frequently may form topics, but frequent words that

are more dispersed may as well. This is very difficult to handle in a quantitative model that reduces all kinds of linguistic aspects to a single dimension. In fact, lexico-quantitatively very similar strings may even have entirely different meanings.[12] Consider for example *The Queen of the Narnia stepped down* and *The Queen of England stepped down*. While nearly identical in phrasing, the two sentences denotate entirely different genres and semantics (ontologies, implications, consequences, semantic fields, and frames). If the corpus further contained *The King of Gondor reached his goals* and *The King of Spain reached his goals*, based on this information alone, any string-similarity-based algorithm would rightfully sort the first two sentences into one topic and the second two into another, while in fact, their distinctivity does not say much about their content. Similarly, the sentences *There was a medical doctor on the plane. He could not help* and *There was a doctor of philosophy on the plane. He could not help* denote different themes and even genres.

2. **Language dynamics are omnipresent and consequential, and topic modeling is unable to account for that.** Language is not stationary. It changes not only over long periods of history, but constantly. Two of the processes and linguistic features that influence the occurrence of words within a document and a corpus are standardization (Ferguson, 1997; Laitinen, 2004; Schmidlin, 2011) and productivity (the coinage of new words, Baayen (1994); Bertram et al. (1999)).
   These do not just lead to quantitative imprecisions, but change the qualitative output of a model. When a new field of study, movement, process, etc. arises in the world, there is at first a proliferance of terms to refer to it. Over time, terms are standardized and differentiated. By then, same or similar concepts are referred to with fewer terms. Some of the previously overlapping terms may become free to be used for other concepts, or just die out. Over time, words are not simply added to the lexicon, but also differentiated in meaning. However, this process is not initiated a single time, but constantly – whenever something new happens, language adapts to be able to describe it (productivity, diversification), and speakers converge on certain mappings of those concepts and words (standardization). This metaprocess overlaps between all newly arising processes in the world – the initial set of terms (pre-new-process) was also the result of developmental processes of diversification and standardization.
   Over time, more and more words exist in the corpus, though not necessarily also in the active vocabulary of each speaker. In a topic model, these would rightfully be sorted into more diverse topics. However, some of these words may in fact refer to the same concepts, while others denote genuinely new concepts, and some may overlap.
   Overall, the tendency will be an overestimation of the diversification of topics over time, i.e., yield an artifact of higher topic diversity. Correspondingly, a diversification of topics over time is in fact a common observation in reported topic models (Wang et al., 2017; Pisarevskaya et al., 2020; Laubichler et al., 2019), albeit without discussion of the influence of language dynamics. It is of course true (and somewhat trivial) that in a world of innovation, new things arise constantly, hence a degree of diversification is to be expected in any time series over discourse. It is still necessary to disentangle the expectable, trivial diversification from the linguistic artifact and an observation relevant to the subfield.

[12] This of course also depends on the type of meaning one refers to, for example intension – the conceptual meaning – vs. extension – the representation in the world – vs. pragmatic meaning or functionality vs. truth-conditional meaning etc. (see Herbelot and Ganesalingam (2013); Herbelot (2013); Lewis and Steedman (2013) for some thoughts specific to distributional and computational semantics). The scope of this paper does not allow to go into more detail.

Topic modeling further presumes that topics are static – the words comprising a topic do not change – and that only the prevalence of the topics changes:

> "Once latent topics are trusted by a variety of means, sociologists can begin to study how they vary over time. In so doing, one can identify the ebb and flow of different language-domains or research-areas within a field", McFarland et al. (2013, 8).

However, in reality, it is obvious that they are not. For example, the topic *computer networks* covered very different concepts in 1990 compared to 2020, and some of the relevant concepts arose and then died out in the meantime. Some would even argue that topics evolve and devolve within a single text or smaller unit:

> "(...) Things have begun perceptibly to change, to wit, the return of interest in thematics, i.e., in capturing the information available but disseminated throughout the text like shifting mists (...). Trying to grasp dispersed information – a moving target without fixed meaning, as indeed topics won't stay in place for the length even of a moderately short sentence – is what thematics is aiming at", Hogenraad et al. (2003, 222).

While corpora can be split by time, modeling changing units over time is a conceptually, namely ontologically, difficult task. Even if models for each year or decade were computed, their interfaces would require definition (which topics in t_1 map to which topics or groups of topics in t_2), and each of those interfaces would suffer from the same uncertainty regarding validity and heuristics as the total model. Text-internal language and topic dynamics are even structurally unaccountable for within a topic model, since topic models are computed from a so-called *bag-of-words-approach*, which loses track of any internal structure.

4.4 Topic prevalence is not an accurate quantification of meaning

One of the major desiderata of text-based research lies in the estimation of how certain aspects of meaning change over time or between factors. It is tempting to view topic prevalence as a representation of "how much of a corpus is about topic x". However, this is a questionable simplification in many ways.

1. **The way topic modeling estimates topic prevalence does not align with the way topics are labeled and perceived.** Human raters a) only consider the most frequent or most distinctive words of a topic, typically the first 10 or less, and b) grasp an underlying concept based on selective perception, they filter irrelevant words in their categorization – a topic that has 9 words clearly belonging to a common theme would likely be categorized as a good topic, even if the 10th word would not fit.

   Topic *prevalence*, i.e., the estimated proportion a topic makes up in a corpus, on the other hand, is computed not only from highly topic-relevant words, but from *all* words sorted into the same topic. This includes a number of conceptually weakly related terms, which still quantitatively contribute to the prevalence of the topic, and some words that are somewhat randomly sorted into the same topic. For example, if a topic model had a topic made up from the set {*news, journalist, TV, station, broadcast, report, article, host, radio, cat*}, a human rater would likely say this is a media topic, and assign the computed prevalence to it. However, the prevalence includes all counts of the word *cat* with only some of the other topic words. If the next five words in the topic – that the human rater might disregard for their labeling, because they may only consider the most frequent or distinctive words for their labeling – were *food, bed, dog, walk, fun*, then the topic prevalence would include parts of documents that have nothing to do with media,

but instead with cat food or dog beds. Changing the number of words considered for the labeling only shifts the problem, because a topic can consist of hundreds of words. Unless one finds a category that fits *all* of the words sorted into a topic, the prevalence estimation is necessarily inaccurate, because it is computed from conflated topics. Thus, while being incomplete with respect to the lexical inventory of the corpus, topic modeling is in fact *overcomplete* in estimating topic prevalence.

2. **The definition and quantification of co-occurrence is a linguistically daunting task, and requires consideration of parameters such as genre or text type and length.** Counting words is hard. It is a common joke among linguists that nobody knows what a word is. Words are quite difficult to define, because they interact with aspects from all linguistic layers (syntactic, morphological, phonetic, semantic, pragmatic, graphematic), but counting words and making sense of the word count in a machine context is known to be notoriously annoying among corpus linguists – so much that, in fact, corpus linguistics usually counts tokens (strings separated by whitespace) to avoid the discussion altogether. For example, if a semantic word consists of two graphematic words (such as *Cold War*), this has repercussions on the quantitative model, although lexicographically it would usually simply be treated as a single word.
   This also bears a conceptual problem – unless all words that are split in writing are accounted for and changed to continuous strings in pre-processing, their combined meaning is lost to the model. Similarly, homographs (words that are written the same but denote different meaning, such as a state *bar*, a *bar* in a bar graph, and a *bar* serving drinks) cannot be distinguished by the model. This has qualitative repercussions, but it also systematically skews the quantitative output of the model.
   Further, similar meaning can come in more or fewer words (*Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo* vs. *buffalo from Buffalo, that are buffaloed by Buffalo from buffalo, in turn buffalo buffalo from Buffalo*).[13] While the occurrence of the string *buffalo* makes up 100% of the first example, it is down to less than 50% in the second example, although the second sentence is merely an explication of the first; and further, technically, not all *buffalo* strings are the same word, since some of them are verbs, some adjectives, and some nouns. For a precise quantitative model, decisions around these issues have to be made, and implications of the method require consideration in the interpretation of the output.
   This also affects the baseline of what counts as co-occurrence. While stylometry and information retrieval typically consider documents, sometimes at book-length, for their analyses, linguistic ideas of co-occurrence are usually concerned with words in somewhat proximal position to one another. This can be defined by the distance within a window of *n* words from one another (typically between 3 and 10), positionally through co-occurrence in the same clause, sentence, paragraph, or chapter, or syntactically through certain relationships such as verb and object. Counting words as co-occurring when they are some 200 pages apart is a stretch of the concept, because it seems unlikely that the reader will make the semantic abstraction between the two in the same way they would in a window of 5 words. This invites unwelcome repercussions on topic abstractions. For example, in this section, the word *cat* will occur several times. Earlier in this paper, the word *taxonomy* occurred a number of times. From close-reading, the reader would not consider this a paper about the taxonomy of cats. However, statistically, relative to the other papers of

[13] https://en.wikipedia.org/wiki/Buffalo_buffalo_Buffalo_buffalo_buffalo_buffalo_Buffalo_buffalo

the author, this is the *cat* + *taxonomy* paper, even though those terms do not even occur in the same section.

The longer a text is, the more likely it is that its words will span various, and often disconnected, topics, but a bag-of-words approach cannot make that distinction. Splitting the text into parts for a topic modeling can be (and has been) done, but that does not resolve the challenge of providing a definition of what constitutes a part of a text. This is conceptually unresolved and unresolvable through linguistic means alone. In scholarly applications of topic modeling, text lengths range from tweets (Chen et al., 2019) to novels (Liu and Jin, 2020) and collections of novels (Tangherlini and Leonard, 2013).

3. **Above-chance word co-occurrences are practically meaningless for most words in a corpus**. This puts the analysis at risk of overinterpreting statistical distinctivity. For any hapax in a corpus, co-occurrence with *any* other word is statistically highly unlikely, and distinctivity can often be reached simply from the fact that most words are rare and can only occur within a limited number of contexts. Due to the large number of possible word combinations, *almost everything* in a corpus is unlikely to co-occur by chance.[14]
4. **Quantifying meaning is conceptually hard.** Topic modeling first extracts co-occurring word bundles, then estimates their prevalence in a corpus. This is typically understood as (a) the words belonging to the same word bundle form a coherent meaning, and (b) their prevalence or proportion in the corpus is expressive of their relevance to the corpus or the writers.

   The implication of (a) is that words that co-occur in certain distinctivity – i.e., co-occur with one another, but not across topics – form more coherent semantic units. The implication of (b) is that, in order to speak about something (a topic), I use the same words, and the more relevant it is to me, the more I use those same words. While both have a certain degree of common sense to them, they are risky in a quantitative analysis: both imply a *scalar* or *gradual* model of topicality and relevance. Since prevalence is given in percentages, another implication is linearity: mathematically, a topic can make up anywhere between 0% and 100% of a corpus, and 2% is twice as relevant as 1%.

   This is a metatheoretical assumption[15] that implies acceptance of the premise that meaning is quantifiable from string matches alone, that it scales linearly (more of the same words mean more of the same meaning). However, extrapolating from the relative frequency of a word is problematic, because the limitations of said frequency are generally unknown. Fig. 1 illustrates this humorously.

   The first problem with this is that words do not behave like that. Since in all corpora, most words are hapaxes, their relative frequency depends more on the size of the corpus (is it one in a million tokens or one in a billion tokens?) than their relevance. Additionally, there are several processes guiding lexical frequency that have little to do with their relevance to a topic, for example burstiness and priming (raised local probability of a word to occur after it has been used, Bock (1986); Hoey (2012); Gries (2005); Madsen et al. (2005); Pierrehumbert (2012)), text structure (some words tend to reliably occur in

[14] The combinatorial potential of words is usually underestimated. For example, Shadrova (2020) calculates the combinatorial potential of the verb and accusative object lexemes as they occur (i.e., the number of verb lexemes that take accusative lexemes times the number of accusative lexemes) in a small corpus containing only 21 relatively short (<1000 tokens) and thematically similar texts to be several magnitudes above the estimated number of atoms in the universe.

[15] "An example of such a metatheoretical assumption could be that the more times a given term appears in a text, the greater is the likelihood that the paper is about the concept that is expressed by that term. Metatheoretic assumptions are thus broader and less specific than theories. They are more or less conscious or unconscious assumptions behind theoretical, empirical, and practical work", Hjørland (1998, 607).

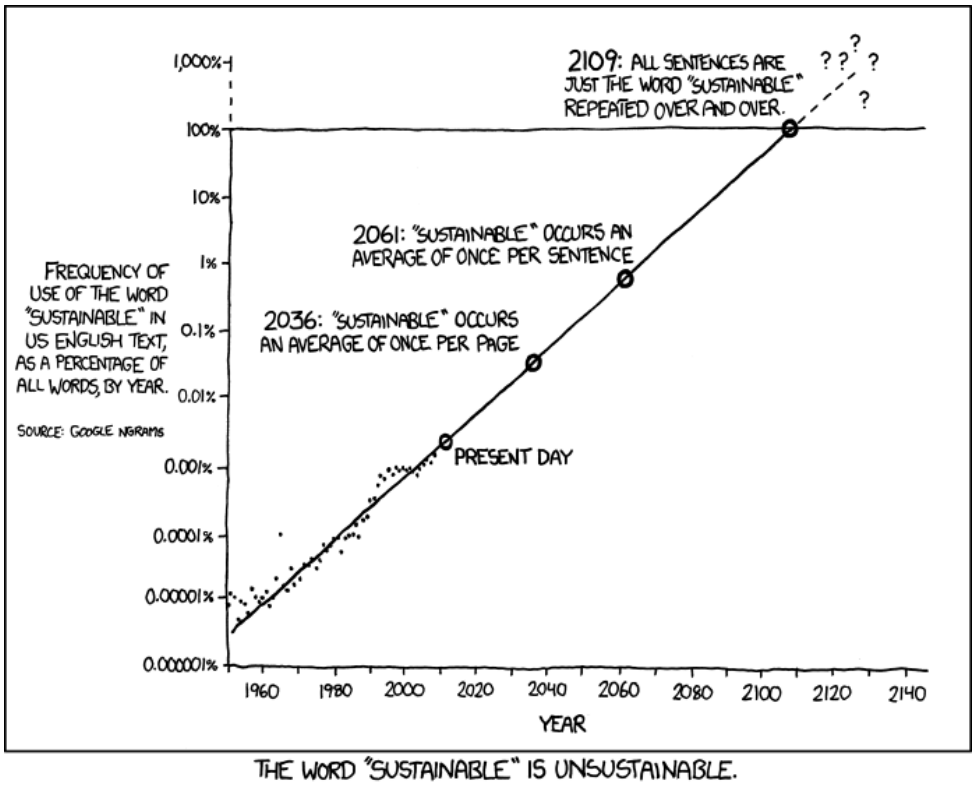

Figure 1: An example of incorrect extrapolation from word frequency in a corpus to real life properties, https://xkcd.com/1007/

certain parts of the text, then disappear, such as *acknowledgement*), and convergence in multi-authored documents (the tendency to converge in the use of syntactic and lexical choices and semantic frames in dialogue, etc., Pickering and Branigan (1998); Steels and Loetzsch (2006); Pardo (2006), and many others. There are some attempts to account for burstiness in topic modeling (Doyle and Elkan, 2009; Madsen et al., 2005), but that does not solve the underlying question: how do we know "how much" of a text is about x? Consider the following examples:

(a) Johan likes cats.
(b) Johan is a cat.
(c) Johan and Katharina like cats.
(d) Johan is a cat, but Katharina is not a cat.
(e) Johan is definitely a cat, but Katharina is not.
(f) Johan is a feline.

How much of each sentence is about cats? In ex. a), a proposition is made about two referents, one of them being a cat. In ex. b), a predication is made – both Johan and the cat refer to the same referent. is b) then "more" about cats, because it does not have another referent? Or is it less about cats, because it is a statement about one specific cat, not the species? Ex. c) is like ex. a), but now with another referent. Quantitatively, the word *cats* is down from 1/3 to 1/5. Is it less about cats than a)? Ex. d) is similar to ex. b), but with another non-cat referent. Quantitatively, it is less about cats than b) (1/4 words in b), 2/10=1/5 in d), even though *cat* occurs twice in d)), and it includes a cat under negation. Does that make it "even less" about cats? How about ex. e), where the first cat is emphasized, the second is deleted? Does the *definitely* make it "more" about cats? Ex. f) does not contain the word *cat*, is it still about cats?

These are not rhetorical questions. In a real-world research scenario, one may not have to weigh simple and similar sentences like these. In practice, however, annotating high-level categories such as *areas of law* or *narrative passages* is notoriously difficult and requires careful development of guidelines. Once the concepts are clear and quantifiable units such as 'sentences mentioning at least one of those concepts' or 'documents that contain at least five mentions' need to be decided on and iteratively validated. This scholarly process it tedious and challenging, but all modeling of meaning is – there is a reason for why the formal modeling of meaning is a subfield of philosophy, linguistics, cognitive science, and AI research each. A more or less precise quantification of meaning may not be impossible to achieve, but can surely not be approximated by counting words in a bag out of context or embedding into a subject-specific framework.

4.5 Topic modeling does not allow for conscious scholarly parameter setting

Topic modeling in LDA allows for a number of parameter settings. Aside from the obvious choice of algorithm, preprocessing and the number of iterations that are performed, these also include the convergence threshold, the number of topics (unless that is defined through preprocessing through NMF or PCA), and hyperparameters α and β which refer to aspects of the distribution.[16] However, neither of those parameters correspond directly to a a number of decisions that are required in the categorization of content. These include at least the following parameters:

- Granularity: How many topics are there? How abstract or concrete can they be, do they all belong to the same level of granularity or is there a taxonomy of some sort?
- Explicitness/implicitness: Are topics explicitly stated and perhaps even syntactically defined (topic: *King/Queen*) or are they implicit (topic: *grief/loss/change/etc.)*?
- Ambiguity: Are topics clearly distinct, can they overlap?
- Scope: How far can topics or themes stretch in the document? How local are they?
- Nomothetic vs. idiothetic vs. idiosyncratic categorization: Are topics defined relative to a full taxonomy or ontology of things (nomothetically), idiographically, i.e., with respect to a single branch of things (for example, *Kafka and estrangement*), or idiosyncratically only for a single leaf within a branch of things (for example *beetle and estrangement*)?

Some discussion on hyperparameter setting in LDA is provided in Asuncion et al. (2012) and Panichella (2021), but the general practice and recommendation is to experiment with topic numbers until one reaches a satisfactory level of semantic coherence or clarity. While acceptable in applied contexts, this is obviously epistemologically problematic, since it invites confirmation bias (stopping at the model that yields results most in line with existing assumptions and ignoring other types of evidence) and apophenia (reading patterns into random data). However, it is also not conceptually the same as the isolation of the parameters listed above – the number of topics is a single dimension, while the parameters exist on several dimensions. The adjustment of topic numbers is an intervention within a dimensionality reduction. It is hard to tell on which of the fused input dimensions it was actually performed.

For example, granularity exists on a continuum – some topics (in life) are more wide-ranging, others are more narrow, and it is easy to construct a topic of intermediate granularity between most granularities of this kind. However, it is plausible to assume that there are some more or less dense ranges on this continuum, i.e., areas where higher or lower topic coherence or a more

[16] $alpha$ and $beta$ reflect constraints on the mixture of topics per document and words for topic, respectively. High α: documents may contain many different topics, not one or two topics specifically; low α: documents should be classified through a small number of topics.

successful taxonomical mapping is expected.[17] However, in a topic modeling, a higher number of topics may correspond with lower granularity, or lower degrees of ambiguity, or more limited locality, or – likely – all of the above. In practice, no mapping of higher or lower granularity by higher or lower topic number seems to be possible. The same is true of ambiguity, scope, or different perspectives such as nomothetic vs. idiothetic groups. It is impossible to consciously search for specific ranges on these diverse dimensions by adjusting topic numbers.

Since implicit meaning *is not included in the words*, but only *read into (or out of) the proximity of words*, no parameter can even theoretically be conjured to define what one is looking for. The risk of apophenia is particularly high here. In order to extrapolate the most useful or knowledgeable topic distribution, it should be possible to perform experiments with granularity, ambiguity, etc., but topic modeling is structurally incapable of this.

4.6 Topic modeling underexplores the analytical potential of the humanities

By relying on frequency of co-occurrence *and* distinctivity in a massively reduced information space, the inner workings of topic modeling streamline content distributions to the most prototypical categorization. This is the best case scenario – in the worst case, they produce somewhat random or even misleading results.[18] Even in the best case scenario, the already streamlined information then runs through the human mind of the researcher, that, in the absence of complex contextual information as it exists in close reading, and without specific methodological guidelines for the interpretation of topic modeling, will further amplify this prototypicality by means of Gestalt perception and cognitive biases.

This has recently been shown empirically by Gillings and Hardie (forthcoming), who ran a study comparing an eyeballing approach based on LDA-extracted topics from news articles to a close-reading approach of the same documents. Out of the twenty topics, they only find one matching topic between the two approaches ("wind farm development in Scotland"). Mismatched topics include major divergences such as "farming and agriculture" (topic modeling) vs. "environmentally-friendly shopping" (close reading) or "community" (topic modeling) vs. "terrorism" (close reading), but also more gradual shifts such as "Labour and Conservative" vs. "Labour leadership race" or "wildlife" vs. "endangered or extinct wildlife".

Rather than synthesizing observation into knowledge under consideration of all complexity, the researcher is limited to connecting dots. Optical illusions, for instance images that can either be perceived as a duck or as a rabbit, work best with just the outline. Adding obvious fur or feather patterns would result in less superposition of the two possible shapes. A topic model can be such an ambiguous outline – however, unlike in the optical illusion, the studied text may not be ambiguous at all. Topics derived from topic models remain constructions interpreted *into*, rather than read *out of* the combination of words present in each topic. But since only little

[17] A similar type of ideal category level is known as *basic-level categories* in perception and learning psychology (Rosch et al., 1976; Hajibayova, 2013; Eimas and Quinn, 1994; Markman and Wisniewski, 1997). A basic-level category is a perceptually salient, easy to detect and memorize object-type category in the world, such as *a car* rather than *a vehicle* or *a Ford*; or *a dog* rather than *a canine* or *a Border Collie*. Basic-level categories are also typically used to refer to newly introduced referents without further context ("I'm by the tree" or "I'm by the oak" rather than "I'm by the Japanese evergreen oak"). They can change with experience, for example, a dog owner may perceive the Border Collie as a basic-level category, while a car salesman might think of a type of Ford as basic level.

[18] This was shown by Schmidt (2012) in a topic modeling-based analysis of references to the whaling industry which implies connections that have never existed. In the case of whaling, this is relatively easily rectifiable, but in more obscure topics, misleading topic modeling output may lead to entirely unfitting models of history or literature.

information is left, especially from a large corpus, the interpretation of the topic model can be like a flipping image. Over time, this underexploration of the existing analytical ability accumulates a lot of insecure and unsynthesized information, even though other ways of knowing do exist and could clarify the issue promptly. Topic modeling effectively cuts the link between the researcher and their actual research object, which is the total collection of text and not its reduction, as Frédéric Clavert has helpfully pointed out. Rather than adding to the stability of the epistemological structure and conceptual clarity of a subject, the uncritical employment of computational methods can effectively work as a limiting factor in the understanding of large text data.

V CLOSING THE GAPS: TOWARDS BEST PRACTICES

While all models are reductionist in nature and abstract from details present in the data, topic modeling does so with reliance on undefined linguistic interfaces and in fairly far-reaching, unpredictable, and so far unmodeled ways.

It is clear that the ideas of a) unshifting semantic spaces and b) meaning detection through string match counts are conceptual simplifications for the sake of operationalization. It is less the fact that they are such simplifications that is epistemologically problematic, and more the unpredictability and density of this dimensionality reduction: it is quite impossible to tell precisely which information has been lost in the condensation from the initial text to bags of words and finally to topic models.

The real challenge in using topic modeling for scholarly purposes then lies in the clarification of the scope of evidence it provides (cf. Leonelli (2019)), in the development of best practices around the argumentative embedding of topic modeling results, and in the formulation of a research program that would result in the methodological integration of topic modeling with linguistic and text-based research to a degree that would allow for its use as an independent research method.

Topic modeling without massive validation does not provide better, more objective, or more exact evidence to most research questions than close reading would. Hence the focus of the argument should be on how the evidence it does provide is in fact relevant to the research question and how its uncertainties affect the argument.

At present, however, it is not common practice to relate topic models to research questions in a detailed fashion. Instead, it is frequently presented as "exploratory evidence" or "a new perspective on data".[19] This in itself is unusual for the empiricist, and it does not appear to be a common case that those studies are then followed up with confirmation, replication, or disconfirmation in later studies. This clashes with the concept of exploration in data, and in any case, it is not common practice to publish data exploration in the quantitative fields. Where it is done, it is within the context of future research agendas or proposals, as incidental findings in the context of other, hypothesis-based analyses, or where resources are presented. It would be highly unusual to publish a mixed-effect model or an ANOVA to an experiment and along with it add all analyses and data wrangling to document the process, or suggest that some of the accidental plots hold high potential for further analysis without linking that back to aspects of the theory. If raw topic modeling output can be taken seriously as a new, interesting, and

[19] See for example Erlin (2018, 3), Carter et al. (2016, 1300–1301), or Rhody (2012, 19): "(...) I suggest that topic modeling poetry works, in part, because of its failures. Somewhere between the literary possibility held in a corpus of thousands of English-language poems and the computational rigor of Latent Dirichlet Allocation (LDA), there is an interpretive space that is as vital as the weaving and unraveling at Penelope's loom".

relevant scholarly exploration of data, than anything can, and no further methodological debate is required.

If, on the other hand, methodological scrutiny is accepted as a requirement for scholarly research, then any contribution requires argumentative embedding. Why and for whom is it interesting, relevant, or new – which research question does it answer and how reliably so?

While the topic model itself reflects an objective reality, namely the result of a maximization of a distribution function over the words in a corpus, the interpretation of a topic model is not objective. It is abductive, which means it is contextualized. This contextualization reflects the frame of reference of the individual scholar. It is of crucial importance to make the underlying contextualization, the model, explicit, both through hypothesis-based work and by tying results back to the theoretical and conceptual debates in the field. Reporting topic modeling results as "naked", "exploratory" results puts them at risk of becoming Troyan horses sneaking into scholarly discourse as confirmed and even somewhat objective *knowledge* rather than the conceptually very basic types of *information* that they really are. This carries problems with reliability, validity, and objectivity. Patterns, even if they may be striking and appear relevant, are not necessarily meaningful and do not classify as epistemes. Their relevance does not stem from being obvious, but from their embedding into a scholarly argument (cf. Dixon (2012)).

Accepting exploratory results from topic models into the research literature also holds the risk of invalidating efforts to cautiously and explicitly model knowledge in the humanities and social sciences through the advent of computational methods. If topic modeling results are viewed as equally convincing or complementary to much more in-depth and more developed analyses in quantitative and qualitative research, a culture of expectation around strong claims may develop more easily. This would, in the long run, harm the humanities and social sciences and their scholarly reputation.

If researchers in the text-based fields insist on using topic modeling as a research technique, a program for its methodological integration needs to be developed. This would need to include at least the following issues:

a) Do topics exist as a definable and quantifiable entity in text?
b) Is there a statistical correlate of topics of certain granularity?
c) How similar are the various topic models of a corpus? Are there better and less good models, and can they be quantitatively determined without reliance on massive human evaluation?
d) How accurate is the topic modeling estimate of topic prevalence compared to the quantification of prevalence based on scholarly modeled topics? How is this affected by text length, genre, and other text-linguistic factors?
e) Which research questions is topic modeling actually suited to answer, outside of "is there at least one statistically distinctive partitioning of a corpus into groups that fulfills a certain requirement"?
f) What are the effects and artifacts of language dynamics on the estimation of topic diversity?
g) Considering the massive reduction of information, how well is topic modeling actually suited for corpus exploration? How does it compare with close reading? How does it compare with other algorithms, such as k-means clustering or PCA? When and where is loss of information an admissible simplification, and where does it turn misleading (Schmidt, 2012)? How do we tell one from the other?

h) How can topic modeling be leveraged in text-based research without overreliance on its shaky results? For example, Wendel et al. (in press) use it as a lexical pre-filter to find words pertaining to areas of law disregarding topic distribution or prevalence, and then construct categories from subject-specific knowledge. Tangherlini and Leonard (2013) suggest three exploration techniques for different use cases in literary studies. How can these and other approaches be synthesized into a well-defined framework for a combined quantitative and qualitative text analysis?

In the absence of such a program, skepticism around topic modeling and other information extraction techniques in text-based research is advised. Best practices should include clarification of the research question and the mapping of categories to subject-specific models, explication of hypotheses and choices in interpretation, a well-argued reason to include the type of evidence provided by the technique and sensitivity to its caveats. The default should be to *not* rely on unembedded topic modeling results unless there is an excellent reason for it in a specific study.[20]

One way to partially avoid the limitations of topic modeling without giving up entirely on the idea of distant reading lies in the simultaneous consideration of several layers of density of complexity, sometimes referred to as multiscale reading (Moa and Ross, 2019). This is implemented for example in the textual exploration software IRaMuTeQ (Camargo and Justo, 2013; Souza et al., 2018), which maintains the link between the concrete reality of the texts involved and its abstractions through statistical classification and other clustering algorithms. Importantly, the focus here is on the *qualitative* analysis. This avoids many of the issues previously discussed, because no unwanted assumptions about scalarity, linearity, or stationarity are implied. However, unless all texts are looked into through both the lenses of close and distant reading, a relatively high degree of uncertainty remains – and if they are, the necessity for distant reading is called into question.

If quantitative analysis of text through measures of information extraction is stipulated as unavoidable in some context, a way to avoid the most problematic aspects of topic modeling lies in the consideration of algorithms that do not rely on bag-of-words approaches and/or topic independence. Lamirel et al. (2020) present such an approach in a feature maximization clustering of words in papers pertaining to the field of Science of Science in China. In feature maximization, words are modeled as features and feature combinations are compared in terms of similarity or distances. This approach, unlike topic modeling, does not assume topic independence, but is instead able to model the relationships between clusters, which can be visualized in so-called contrast graphs. This allows for some understanding of cluster granularity and connectivity, and since the algorithm is entirely unsupervised and does not rely on hyperparameter setting, lends itself less to unintentional data tweaking.

Clustering approaches of this kind provide clear advantages in terms of the alignment between the conceptual and mathematical model in some ways, and they may yield better results in terms of the quality of the extracted content. Lamirel et al. (2020) show this in a comparison with LDA-extracted topics of the same corpus. However, clustering does not solve the underlying linguistic issues linked to the interpretation of meanings of word co-occurrences outside of immediate context. This means that many of the pitfalls discussed previously remain – including the fact that there is no linguistic model that maps word co-occurrences or their frequency to

[20] One such example is the study by Block and Newman (2011). It shows a persisting diversity of topics as counterevidence to the claim that history journals limit their scope to women's history topics. For this specific research question, simply the evidence of diversity suffices to counter the claim, but this is not typical of most research.

specificable units of meaning such as topics. While in their study, Lamirel and colleagues find high convergence between the extracted clusters with expert descriptions of the history of Science of Science in China, this cannot be seen as definitive proof of the algorithm's capability to uncover objective realities underlying text. The word clusters are still subject to interpretation, they do not naturally correspond to certain topics or fields, but are constructed as such through an expert, which in a post-hoc interpretation can be affected by apophenia and confirmation bias.

In order to be certain that the data supports the hypothesis, word groups that would be interpretable as indicators of a certain topic would have to be described prior to the analysis. In Lamirel et al. this may not be the case because the expert (a professor of the field) would also be highly familiar with the included texts, thus not actually relying on distant reading – in a biographical sense, this can even be viewed as multiscale reading. Unlike this, distant reading is frequently proposed as an *alternative* to the close reading of large collections of texts, and in fact as the only way to approach text in the era of big data. A specific research question may justify the employment of specific techniques of information extraction such as feature maximization clustering. But generally speaking, the degree of epistemological and linguistic uncertainty remains high wherever meaning extraction relies almost exclusively on words and not on deeper annotations describing higher-order structures of meaning or ontologies relevant to a given field.

VI CONCLUSION

The discussion in this paper has shown that the evidence provided by topic modeling is conceptually weak in a number of ways. Even for a statistically optimal model, the interface with linguistics is largely undefined. Since validation from within is largely impossible, evidence from topic modeling only has very limited weight on its own.

Wherever statistics, combinatorics, and complexity come into play, common-sensical explanations rarely suffice. It is easy to tell a whale from a parrot, but correctly identifying different subspecies of parrots is much more difficult. In the same way, it is obvious that in a corpus in which every other text mentions the word *judicial*, compared to a corpus in which every other text mentions the word *rabbit food* and does not mention *judicial*, different topics in any sense of the word are likely to be at play. However, this is not equivalent to a *scalar* or *gradual* model of aboutness from statistical features of text.

Frequency of (co-)occurrence is not a sufficient marker for category boundaries. This is due to varying granularities of topics, but more so due to the specifics of lexeme distributions. The assumption of unchangeability of frequency of (co-)occurrence per topic and the maximization of discreteness between topics can be further problematic, because it biases topic models to the most prototypical patterns and systematically counteracts the recognition of finer-grained distinction and dynamic or shifting boundaries. This leaves the model structurally incapable of exhausting the analytical potential of the social sciences or humanities. It also creates an impression of objectivity ("this agrees with a general consensus in the field") and obvious coherence ("these words are obviously connected"), which may rely more on the human bias towards prototype perception, apophenia, and confirmation bias than 'actual', meaningful coherence.

The problem with topic modeling is not that it has no potential of yielding coherent or even useful results, but that those results are reductionist in relatively unpredictable ways and that there is no obvious way to integrate such results as subject-specific knowledge without a high degree of methodological and theoretical effort. To accept its results blindly is to reify them

as knowledge where they are, in fact, merely a largely unmodeled type of information. If topic modeling and other probabilistic text mining techniques are to be taken seriously in the advancement of knowledge in the social sciences and digital humanities, they must face the same rigid analysis and review as other methods.

This also highlights the importance of clear research questions and a systematic approach towards the choice, application, and validation of methods. To define the scope of evidence of a method is to clarify the research questions it is confidently able to answer relative to a type of data: "With this method, question X can be answered provided there is access to data of type Y, and the answer can be integrated into a specified theoretical model in ways a, b, or c".

Researchers in the natural sciences have a toolbox of largely well-defined methods to use for clearly defined purposes – they know when to use a pH-test and when to use a tachometer, and they would likely not measure and report the acidity of a fluid or the speed of a rotating object for exploratory purposes, unless they ran a series of tests that included acidity or rotation speed among other aspects. This methodological confidence did not appear over night, but developed over the course of centuries of scientification. Similar efforts towards systematization are necessary in the computationally oriented text-based research fields.

Without such embedding, the inclusion of spurious or random information through topic modeling is nearly unavoidable. While that may matter little in applied contexts, because spurious results can be filtered out in the application feedback loop, the same is not true of the social sciences or digital humanities. If we were to trust the machine, we would be forced to accept whatever latent dimension is detected as objectively there – and consequentially, if we were to take standing research seriously, we would have to consider the evidence provided by the machine in all future research. Blindly accepting machine-generated topic distributions into our understanding of the subject is similar to equipping libraries with books that have semantically coherent titles (as determined by human raters and the machine) but mostly blackened pages and referring to this library as the library of our knowledge of the subject.

By themselves, topic models do not provide better, more objective, or more exact evidence than other types of category construction from data. Topics are not a uniquely defined concept, and cannot be uniquely defined for any text. They are, in fact, not even a linguistic concept per se, but are constructed from the words of a text, its context, and even the space between the words of a text. The decision of whether the words in a topic model topic do in fact correspond meaningfully to an acknowledged or a plausible new category in their discipline remains, and must remain, with the scholar. If topics, themes, etc. are intended as scholarly categories, they require a process of debate to be constructed and synthesized, thereby creating scholarly knowledge.

A wide range of methodological literature shows that high standards are set upon the clarification of argumentative logic and category construction in the qualitative social sciences and humanities. The same should apply to a) quantitative text-based research in general and b) computer-assisted methods in particular.

References

Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., and Zhu, M. (2013). A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning*, pages 280–288.

Arora, S., Ge, R., and Moitra, A. (2014). New algorithms for learning incoherent and overcomplete dictionaries. In *Conference on Learning Theory*, pages 779–806.

Ascari, M. (2014). The dangers of distant reading: Reassessing Moretti's approach to literary genres. *Genre: Forms of Discourse and Culture*, 47(1):1–19.

Asgari, E., Ghassemi, M., and Finlayson, M. A. (2013). Confirming the themes and interpretive unity of Ghazal poetry using topic models. In *Neural Information Processing Systems (NIPS) Workshop for Topic Models*.

Asuncion, A., Welling, M., Smyth, P., and Teh, Y. W. (2012). On smoothing and inference for topic models. *arXiv preprint arXiv:1205.2662*.

Asuncion, H. U., Asuncion, A. U., and Taylor, R. N. (2010). Software traceability with topic modeling. In *2010 ACM/IEEE 32nd International Conference on Software Engineering*, volume 1, pages 95–104. IEEE.

Baayen, R. H. (1994). Productivity in language production. *Language and Cognitive Processes*, 9(3):447–469.

Baayen, R. H. (2002). *Word frequency distributions*, volume 18. Springer Science & Business Media.

Bakshi, R. K., Kaur, N., Kaur, R., and Kaur, G. (2016). Opinion mining and sentiment analysis. In *2016 3rd international conference on computing for sustainable global development (INDIACom)*, pages 452–455. IEEE.

Bao, S., Xu, S., Zhang, L., Yan, R., Su, Z., Han, D., and Yu, Y. (2009). Joint emotion-topic modeling for social affective text mining. In *2009 Ninth IEEE International Conference on Data Mining*, pages 699–704. IEEE.

Baroni, M., Bernardi, R., Zamparelli, R., et al. (2014). Frege in space: A program for compositional distributional semantics. *Linguistic Issues in language technology*, 9(6):5–110.

Bertalan, V. G. and Ruiz, E. E. S. (2019). Using topic modeling to find main discussion topics in brazilian political websites. In *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web*, pages 245–248.

Bertram, R., Laine, M., and Karvinen, K. (1999). The interplay of word formation type, affixal homonymy, and productivity in lexical processing: Evidence from a morphologically rich language. *Journal of psycholinguistic research*, 28(3):213–226.

Bhatia, S., Lau, J. H., and Baldwin, T. (2018). Topic intrusion for automatic topic model evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 844–849.

Blei, D., Carin, L., and Dunson, D. (2010). Probabilistic topic models. *IEEE signal processing magazine*, 27(6):55–65.

Blei, D. and Lafferty, J. (2006). Correlated topic models. *Advances in neural information processing systems*, 18:147.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Block, S. and Newman, D. (2011). What, where, when, and sometimes why: Data mining two decades of women's history abstracts. *Journal of Women's History*, 23(1):81–109.

Bock, J. K. (1986). Meaning, sound, and syntax: Lexical priming in sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12(4):575.

Boyd-Graber, J., Hu, Y., Mimno, D., et al. (2017). Applications of topic models. *Foundations and Trends® in Information Retrieval*, 11(2-3):143–296.

Boyd-Graber, J., Mimno, D., and Newman, D. (2014). Care and feeding of topic models: Problems, diagnostics, and improvements. *Handbook of mixed membership models and their applications*, 225255.

Bruni, E., Tran, N.-K., and Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Camargo, B. V. and Justo, A. M. (2013). IRAMUTEQ: a free software for analysis of textual data. *Temas em psicologia*, 21(2):513–518.

Carter, D. J., Brown, J., and Rahmani, A. (2016). Reading the high court at a distance: Topic modelling the legal subject matter and judicial activity of the high court of australia, 1903-2015. *UNSWLJ*, 39:1300–1354.

Chandelier, M., Steuckardt, A., Mathevet, R., Diwersy, S., and Gimenez, O. (2018). Content analysis of newspaper coverage of wolf recolonization in France using structural topic modeling. *Biological conservation*, 220:254–261.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296.

Chen, X., Zou, D., Cheng, G., and Xie, H. (2020). Detecting latent topics and trends in educational technologies

over four decades using structural topic modeling: A retrospective of all volumes of computer & education. *Computers & Education*, page 103855.

Chen, Y., Zhang, H., Liu, R., Ye, Z., and Lin, J. (2019). Experimental explorations on short text topic mining between LDA and NMF based Schemes. *Knowledge-Based Systems*, 163:1 – 13.

Chuang, J., Gupta, S., Manning, C., and Heer, J. (2013). Topic model diagnostics: Assessing domain relevance via topical alignment. In *International conference on machine learning*, pages 612–620.

Chuang, J., Roberts, M. E., Stewart, B. M., Weiss, R., Tingley, D., Grimmer, J., and Heer, J. (2015). TopicCheck: Interactive alignment for assessing topic model stability. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–184.

Dahllöf, M. and Berglund, K. (2019). Faces, fights, and families: Topic modeling and gendered themes in two corpora of Swedish prose fiction. In *DHN 2019, 4th Digital Humanities in the Nordic Countries 2019, University of Copenhagen, Copenhagen, Denmark, March 6–8, 2019*, pages 92–111.

de Paulo Faleiros, T. and de Andrade Lopes, A. (2016). On the equivalence between algorithms for Non-negative Matrix Factorization and Latent Dirichlet Allocation. In *ESANN*.

Debowski, L. (2018). Is natural language a perigraphic process? The theorem about facts and words revisited. *Entropy*, 20:85.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

DiMaggio, P., Nag, M., and Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. *Poetics*, 41(6):570–606.

Ding, C., Li, T., and Peng, W. (2006). Nonnegative matrix factorization and probabilistic latent semantic indexing: Equivalence chi-square statistic, and a hybrid method. In *AAAI*, volume 42, pages 137–43.

Dixon, D. (2012). Analysis tool or research methodology: Is there an epistemology for patterns? In *Understanding digital humanities*, pages 191–209. Springer.

Doyle, G. and Elkan, C. (2009). Accounting for burstiness in topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 281–288, New York, NY, USA. Association for Computing Machinery.

Drucker, J. (2017). Why distant reading isn't. *PMLA*, 132(3):628–635.

Eimas, P. D. and Quinn, P. C. (1994). Studies on the formation of perceptually based basic-level categories in young infants. *Child development*, 65(3):903–917.

Erlin, M. (2018). Topic modeling, epistemology, and the English and German novel. *SocArXiv*.

Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. *arXiv preprint arXiv:1909.00512*.

Fabre, C. and Lenci, A. (2015). Distributional Semantics Today: Introduction to the special issue. *Revue TAL, ATALA (Association pour le Traitement Automatique des Langues), Sémantique distributionnelle*, 56(2):7–20.

Ferguson, C. A. (1997). Standardization as a form of language spread. In *Structuralist studies in Arabic linguistics*, pages 69–80. Brill.

Fried, D. and Duh, K. (2014). Incorporating both distributional and relational semantics in word representations. *arXiv preprint arXiv:1412.4369*.

Gillings, M. and Hardie, A. (forthcoming). The interpretation of topic models for scholarly analysis: an evaluation and critique of current practice.

Girolami, M. and Kabán, A. (2003). On an equivalence between PLSI and LDA. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 433–434.

Goldstone, A. and Underwood, T. (2014). The quiet transformations of literary studies: What thirteen thousand scholars could tell us. *New Literary History*, 45(3):359–384.

Goutsos, D. (1997). *Modeling Discourse Topic: sequential relations and strategies in expository text*, volume 59. Greenwood Publishing Group.

Graneheim, U. H. and Lundman, B. (2004). Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. *Nurse education today*, 24(2):105–112.

Greene, D. and Cross, J. P. (2016). Exploring the political agenda of the european parliament using a dynamic topic modeling approach. *arXiv preprint arXiv:1607.03055*.

Gries, S. T. (2005). Syntactic priming: A corpus-based approach. *Journal of psycholinguistic research*, 34(4):365–399.

Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in

Senate press releases. *Political Analysis*, 18(1):1–35.
Hajibayova, L. (2013). Basic-level categories: A review. *Journal of Information Science*, 39(5):676–687.
Hasan, R. (1967). Linguistics and the study of literary texts. *Etudes de linguistique appliquee*, 5:106–121.
Herbelot, A. (2013). What is in a text, what isn't, and what this has to do with lexical semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Short Papers*, pages 321–327.
Herbelot, A. and Ganesalingam, M. (2013). Measuring semantic content in distributional vectors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 440–445.
Hjørland, B. (1998). Theory and metatheory of information science: a new interpretation. *Journal of documentation*.
Hoey, M. (2012). *Lexical priming: A new theory of words and language*.
Hogenraad, R., McKenzie, D. P., and Péladeau, N. (2003). Force and influence in content analysis: The production of new social knowledge. *Quality and Quantity*, 37(3):221–238.
Jacobs, A. M. (2018). The Gutenberg English Poetry Corpus: exemplary quantitative narrative analyses. *Frontiers in Digital Humanities*, 5:5.
Jänicke, S., Franzini, G., Cheema, M. F., and Scheuermann, G. (2015). On close and distant reading in Digital Humanities: A survey and future challenges. In *EuroVis (STARs)*, pages 83–103.
Jockers, M. L. and Mimno, D. (2013). Significant themes in 19th-century literature. *Poetics*, 41(6):750–769.
Kehler, A. (2004). Discourse topics, sentence topics, and coherence. *Theoretical Linguistics*, 30(2-3):227–240.
Kimberly, A. N. (2011). *The Content Analysis Guidebook, 2nd edition*. SAGE PUBLICATIONS Incorporated.
Korfiatis, N., Stamolampros, P., Kourouthanassis, P., and Sagiadinos, V. (2019). Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews. *Expert Systems with Applications*, 116:472–486.
Krippendorff, K. (2004a). *Content Analysis: An Introduction to Its Methodology (second edition)*. Sage Publications.
Krippendorff, K. (2004b). Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433.
Kuang, D., Choo, J., and Park, H. (2015). Nonnegative matrix factorization for interactive topic modeling and document clustering. In *Partitional Clustering Algorithms*, pages 215–243. Springer.
Kuhn, K. D. (2018). Using structural topic modeling to identify latent topics and trends in aviation incident reports. *Transportation Research Part C: Emerging Technologies*, 87:105–122.
Laitinen, L. (2004). Grammaticalization and standardization. *Typological Studies in Language*, 59:247–262.
Lamirel, J.-C., Chen, Y., Cuxac, P., Al Shehabi, S., Dugué, N., and Liu, Z. (2020). An overview of the history of Science of Science in China based on the use of bibliographic and citation data: a new method of analysis based on clustering with feature maximization and contrast graphs. *Scientometrics*, 125:2971–2999.
Lampach, N. and Dyevre, A. (2018). Issue attention on international courts: A text-mining approach. *Available at SSRN 3251186*.
Lau, J. H., Newman, D., and Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.
Laubichler, M. D., Maienschein, J., and Renn, J. (2019). Computational history of knowledge: Challenges and opportunities. *Isis*, 110(3):502–512.
Law, D. S. (2016). Constitutional archetypes. *Tex. L. Rev.*, 95:153.
Lengerich, B. J., Maas, A. L., and Potts, C. (2017). Retrofitting distributional embeddings to knowledge graphs with functional relations. *arXiv preprint arXiv:1708.00112*.
Leonelli, S. (2019). What distinguishes data from models? *European journal for philosophy of science*, 9(2):22.
Levy, O. and Goldberg, Y. (2014). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308.
Lewis, M. and Steedman, M. (2013). Combined distributional and logical semantics. *Transactions of the Association for Computational Linguistics*, 1:179–192.
Lindstedt, N. C. (2019). Structural topic modeling for social scientists: A brief case study with social movement studies literature, 2005–2017. *Social Currents*, 6(4):307–318.
Liu, J., Weinert, A., and Amin, S. (2019). Semantic analysis of traffic camera data: Topic signal extraction and anomalous event detection. *arXiv preprint arXiv:1905.07332*.
Liu, L., Tang, L., Dong, W., Yao, S., and Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5(1):1608.
Liu, X. and Jin, M. (2020). Classification analysis of Kouji Uno's novels using topic model. *Behaviormetrika*, 47(1):189–212.

Liu, Y., Liu, Z., Chua, T.-S., and Sun, M. (2015). Topical word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Livermore, M. A., Riddell, A., and Rockmore, D. (2016). Agenda formation and the US supreme court: A topic model approach. *Arizona Law Review*, 1(2).

Macnamara, J. R. et al. (2005). Media content analysis: Its uses, benefits and best practice methodology. *Asia Pacific Public Relations Journal*, 6(1):1.

Madsen, R. E., Kauchak, D., and Elkan, C. (2005). Modeling word burstiness using the Dirichlet distribution. In *Proceedings of the 22nd international conference on Machine learning*, pages 545–552.

Markman, A. B. and Wisniewski, E. J. (1997). Similar and different: The differentiation of basic-level categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(1):54.

Mayring, P. (2004). Qualitative content analysis. In von Kardorff, E., Flick, U., and Steinke, I., editors, *A Companion to Qualitative Research*, pages 159–176.

McFarland, D. A., Ramage, D., Chuang, J., Heer, J., Manning, C. D., and Jurafsky, D. (2013). Differentiating language usage through topic models. *Poetics*, 41(6):607 – 625. Topic Models and the Cultural Sciences.

Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

Mitrofanova, O. and Sedova, A. (2017). Topic modeling in parallel and comparable fiction texts (the case study of English and Russian prose). In *Proceedings of the International Conference IMS-2017*, pages 175–180.

Moa, B. and Ross, S. (2019). Big data analytics for multiscale reading. In *Doing More Digital Humanities*, pages 199–236. Routledge.

Moretti, F. (2013). *Distant reading*. Verso Books.

Möllers, C., Shadrova, A., and Wendel, L. (2021). BVerfGE-Korpus, v.1.0. https://doi.org/10.5281/zenodo.4551408.

Navarro-Colorado, B. (2018). On poetic topic modeling: extracting themes and motifs from a corpus of spanish poetry. *Frontiers in Digital Humanities*, 5:15.

Panagis, Y., Christensen, M. L., and Sadl, U. (2016). On top of topics: Leveraging topic modeling to study the dynamic case-law of international courts. In *JURIX*, pages 161–166.

Panichella, A. (2021). A systematic comparison of search-based approaches for LDA hyperparameter tuning. *Information and Software Technology*, 130:106411.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, 119(4):2382–2393.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Pettersson, B. (2002). Seven trends in recent thematics and a case study. In Louwerse, M. and van Peer, W., editors, *Thematics. Interdisciplinary Studies*, pages 237–252. John Benjamins Publishing.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21(5):1112–1130.

Pickering, M. J. and Branigan, H. P. (1998). The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and language*, 39(4):633–651.

Pierrehumbert, J. B. (2012). Burstiness of verbs and derived nouns. In *Shall We Play the Festschrift Game?*, pages 99–115. Springer.

Pisarevskaya, A., Levy, N., Scholten, P., and Jansen, J. (2020). Mapping migration studies: An empirical analysis of the coming of age of a research field. *Migration Studies*, 8(3):455–481.

Ramage, D., Manning, C. D., and Dumais, S. (2011). Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 457–465.

Ramirez, E. H., Brena, R., Magatti, D., and Stella, F. (2012). Topic model validation. *Neurocomputing*, 76(1):125–133.

Reber, U. (2019). Overcoming language barriers: Assessing the potential of machine translation and topic modeling for the comparative analysis of multilingual text corpora. *Communication methods and measures*, 13(2):102–125.

Rhody, L. M. (2012). Topic modeling and figurative language. *Journal of Digital Humanities*, 2(1):19–35.

Rice, D. R. (2017). Issue divisions and US Supreme Court decision making. *The Journal of Politics*, 79(1):210–222.

Rimmon-Kenan, S. (1995). What is theme and how do we get at it? In Bremond, C., Landy, J., and Pavel, T., editors, *Thematics: New Approaches*, pages 9–20. SUNY Press.

Roberts, M. E., Stewart, B. M., and Tingley, D. (2019). stm: An R package for structural topic models. *Journal of Statistical Software*, 91(2):1–40.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., and Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., and Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology*, 8(3):382–439.

Schlobinski, P. and Schütze-Coburn, S. (1992). On the topic of topic and topic continuity. *Linguistics*, 30(1):89–122.

Schmidlin, R. (2011). *Die Vielfalt des Deutschen: Standard und Variation: Gebrauch, Einschätzung und Kodifizierung einer plurizentrischen Sprache*, volume 106. Walter de Gruyter.

Schmidt, B. M. (2012). Words alone: Dismantling topic models in the humanities. *Journal of Digital Humanities*, 2(1):49–65.

Schöch, C. (2017). Topic modeling genre: An exploration of French classical and enlightenment drama. *DHQ: Digital Humanities Quarterly*, 11(2).

Shadrova, A. (2020). *Measuring coselectional constraint in learner corpora: A graph-based approach*. PhD thesis, Humboldt-Universität zu Berlin.

Shadrova, A. (in press). It may be in the structure, not the combinations: graph metrics as an alternative to statistical measures in corpus-linguistic research. In *Graph Technologies in the Humanities 2020*.

Si, J., Li, Q., Qian, T., and Deng, X. (2014). Users' interest grouping from online reviews based on topic frequency and order. *World Wide Web*, 17(6):1321–1342.

Sieg, C. (2019). Topic Modeling von Fallgeschichten. *Zeitschrift für Literaturwissenschaft und Linguistik*, 49(4):653–671.

Sollors, W. (2002). Thematics today. In Louwerse, M. and van Peer, W., editors, *Thematics: Interdisciplinary Studies*, volume 3, pages 217–235. John Benjamins Publishing.

Souza, M. A. R. d., Wall, M. L., Thuler, A. C. d. M. C., Lowen, I. M. V., and Peres, A. M. (2018). The use of IRAMUTEQ software for data analysis in qualitative research. *Revista da Escola de Enfermagem da USP*, 52.

Speer, R. and Lowry-Duda, J. (2017). Conceptnet at semeval-2017 task 2: Extending word embeddings with multilingual relational knowledge. *arXiv preprint arXiv:1704.03560*.

Steels, L. and Loetzsch, M. (2006). Perspective alignment in spatial language. *arXiv preprint cs/0605012*.

Sterling, J., Jost, J. T., and Hardin, C. D. (2019). Liberal and conservative representations of the good society: A (social) structural topic modeling approach. *Sage Open*, 9(2):2158244019846211.

Tangherlini, T. R. and Leonard, P. (2013). Trawling in the sea of the Great Unread: Sub-corpus topic modeling and Humanities research. *Poetics*, 41(6):725–749.

Thoma, S., Rettinger, A., and Both, F. (2017). Towards holistic concept representations: Embedding relational knowledge, visual attributes, and distributional word semantics. In *International Semantic Web Conference*, pages 694–710. Springer.

Underwood, T. (2014). Theorizing research practices we forgot to theorize twenty years ago. *Representations*, 127(1):64–72.

Underwood, T. (2017). A genealogy of distant reading. *DHQ: Digital Humanities Quarterly*, 11(2).

van Der Hooft, J. J. J., Wandy, J., Barrett, M. P., Burgess, K. E., and Rogers, S. (2016). Topic modeling for untargeted substructure exploration in metabolomics. *Proceedings of the National Academy of Sciences*, 113(48):13738–13743.

Van Peer, W. (2002). Where do literary themes come from? In Louwerse, M. and van Peer, W., editors, *Thematics: Interdisciplinary Studies*, volume 3, pages 253–263. John Benjamins Publishing.

Wang, C. and Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 448–456.

Wang, Y., Bowers, A. J., and Fikis, D. J. (2017). Automated text data mining analysis of five decades of educational leadership research literature: Probabilistic topic modeling of EAQ articles from 1965 to 2014. *Educational administration quarterly*, 53(2):289–323.

Weitin, T. and Herget, K. (2017). Falkentopics. *Zeitschrift für Literaturwissenschaft und Linguistik*, 47(1):29–48.

Wendel, L., Shadrova, A., and Tischbirek, A. (in press). From modeled topics to areas of law: A comparative analysis of types of proceedings in the German Federal Constitutional Court. *German Law Journal*.

Wesslen, R. (2018). Computer-assisted text analysis for social science: Topic models and beyond. *arXiv preprint arXiv:1803.11045*.

Yang, T.-I., Torget, A., and Mihalcea, R. (2011). Topic modeling on historical newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 96–104.

Young, D. T. (2012). How do you measure a constitutional moment: Using algorithmic topic modeling to evaluate

Bruce Ackerman's Theory of Constitutional Change. *Yale LJ*, 122:1990.

Young, R. L. (2019). *The language and rhetoric of affirmative action: a structural topic model analysis of supreme court amicus briefs*. PhD thesis, University of Iowa.

Yu, L., Zhang, C., Shao, Y., and Cui, B. (2017). LDA*: a robust and large-scale topic modeling system. *Proceedings of the VLDB Endowment*, 10(11):1406–1417.

Zhong, Y., Zhu, Q., and Zhang, L. (2015). Scene classification based on the multifeature fusion probabilistic topic model for high spatial resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 53(11):6207–6222.

Zipf, G. K. (1965). *The Psycho-Biology of Language. An Introduction to Dynamic Philology. 1935*. Cambridge, Mass: The MIT Press.