Foucault's archeological discourse analysis with digital methodology—Discourse on women prior to the first wave women's movement

Heidi Karlsen
University of Oslo, Oslo, Norway and BI Norwegian Business School, Oslo, Norway

Abstract

In *L'archéologie du savoir* (1969), Foucault specifies the method of discourse analysis as the identification of statements. Few digital humanities projects have set out to conduct discourse analysis in this tradition. A certain degree of confusion relates to what the Foucauldian statement consists of and how it can be operationalized in order to identify discourse, using digital methods. This article demonstrates, however, that such an operationalization becomes possible if we take Foucault's own distinction between *statement* (énoncé) and *enunciation* (énonciation) into account. As an example, the article shows how discourse on women prior to the first wave women's movement in Western countries has been identified in almost 7,000 books digitized by the Norwegian National Library. This article concludes that a digitized Foucauldian discourse analysis is possible, using a combination of digital methodology and close reading.

Correspondence:
Heidi Karlsen, Niels Henrik Abels vei 36, Oslo, Norway, University of Oslo.
E-mail:
heidi.karlsen@ub.uio.no

1 Introduction

More than fifty years have passed since Michel Foucault wrote *L'archéologie du savoir* in which he describes discourse analysis as a method. It is safe to say that discourse analysis is an established methodology within the humanities today, as well as in social and educational sciences. This is no less true within the digital humanities; extensive research has been carried out using digital methodology in order to identify discourse. Few digital humanities projects have, however, set out to conduct discourse analysis in line with Foucault's description in *L'archéologie du savoir*. In this work, he specifies the method of discourse analysis as the identification of statements. A certain degree of confusion relates to what the Foucauldian statement consists of and how it can be operationalized in order to identify discourse, using digital methods. This explains why Foucauldian discourse analysis is rare in digital humanities projects.

A Foucauldian statement is an epistemological object, identified by the researcher within a defined archive. In order to identify statements, one observes regularities in ways of speaking, the functions certain ways of speaking have, which depend on the relation of a statement to other statements within a selected material (Foucault, 2014, pp. 109–45). Statements are thus context-dependent and one and the same statement does not necessarily come across through only one combination of linguistic signs. This is one reason why scholars have

https://doi.org/10.1093/llc/fqac022

argued that it is difficult to operationalize Foucault's concept in order to use digital methodology to conduct discourse analysis in his tradition (Erb *et al.*, 2016). I will show, however, that it becomes possible if we use digital methodology to capture, not statements (énoncés) in themselves, but enunciations (énonciations) in Foucault's methodology. Enunciations are unique: someone said something at a specific point in time. The analytical work consists in identifying—in the 'mass of said things' (Foucault, 1969) or assemblage of enunciations—repeated patterns in ways of speaking. We specify these patterns as the far smaller number of *statements* of which we can say that these *enunciations* are concrete actualizations.

Without at all undermining the complexity of Foucault's 'archeological' method, in this analysis I argue that digital methodology, more specifically sub-corpus topic modelling in combination with a bag of word tool and close reading, can be used to conduct Foucauldian discourse analysis as detailed in *L'archéologie du savoir*.

There are several reasons why Foucauldian discourse analysis is still a useful methodology and why combining it with digital methodology is promising. Foucault's 'archeological' method enables the identification of ways of speaking and their functions across different fields of knowledge. Foucault's concept of the archive is crucial for identifying discourse in his tradition. This concept has often been omitted in the use of his method. Today's massive collections of digitized material provide us with new opportunities for creating and analysing archives as 'masses of said things' across disciplines. This might lead to new insight into subjects that have not been much researched or researched within only specific disciplines. As an example I will show how I have used a digitized Foucauldian discourse analysis to identify ways of speaking about women in books digitized by the Norwegian National Library from the nineteenth century, prior to what is often referred to as the first wave women's movement in Western countries.

After a discussion of key concepts in Foucault's archeological method and an overview of previous research on Foucault's method in combination with digital methodology, I show how his method can be operationalized for digital collections of documents and tools for 'mining' them. In the last part of the article, I present the example of a discourse on women that this methodology has made me identify.

2 The Foucauldian Archive

In Foucault's archeological method, *archive* is a key a concept. On the back cover of *L'archéologie du savoir*, we can read the following: 'The domain of said things, that's what we call the archive: which the archeology is destined to analyse.'[1] The concept of the archive is indispensable for Foucault's development of discourse analysis as a method (Eliassen, 2016, p. 74). In a time when digitization processes transform the materiality and accessibility of archives and collections of documents, our research opportunities are altered. It is highly relevant to question how Foucault's 'archival' method and notion of the archive may relate to the research practices that become possible with the digitized documents that are increasingly available to us.

The Foucauldian archive, however, is not simply a collection of documents, records of materials, or the places where these are stored. The archive also designates 'the mass of what has been said' during a specific and delimited period (Foucault, 1969). An 'archive' in this sense is not an object that exists prior to analysis. It is instead an epistemological object for analysis established with the help of a set of criteria. By 'mass of said things,' we must understand the great variety of texts and other 'said things' produced during a specific time period. In order to prepare an archive of a period, even 'everyday' documents, perhaps uncanonized and forgotten, will thus be important to collect. Foucault's concept of the archive as an epistemological object refers to a researcher's gathering of a set of material and the discursive statements she identifies within it.

3 The Foucauldian Statement

In the article 'Possibility of Discourse Analysis Using Topic Modeling', Wonkwang Jo (2019) discusses relations between topic modelling and Foucauldian discourse analysis. He points out that discourse is not 'language activity' in itself and that a topic is not an 'accurate measure of a discourse', but a 'reference for identifying discourse'. Jo provides some interesting reflections in terms of how topic modelling can be a

useful approach to identify discourse. However, he neither mentions the Foucauldian statement nor discusses in detail how precisely the topics can be used to identify discourse as developed by Foucault in *L'archéologie de savoir*. Cecilie H. Sam (2019), in contrast, addresses the Foucauldian statement as well as the concept of the archive. Sam has used the Sifter application to select around 6,000 tweets. These tweets form her corpus, which she also labels her Foucauldian archive. It seems that the article considers this totality of tweets as equally many discursive statements. As Sam herself declares earlier in the same article, however, a statement for Foucault is more a function than a linguistic unit and does not exist in isolation. Statements are thus identified and described based on patterns of regularity in the archive in question. They will be far fewer than the entire material that makes up the archive. Thus, this article does not bring us closer to how digital methodology can be used to identify discursive statements in the Foucauldian sense. Still it enlightens its subject (educational policy and more specifically the Common Core State Standards Initiative in the USA) and brings forth relevant perspectives on Foucauldian discourse analysis. Sam analyses contemporary discourse and one and only one mode of expression (tweets), however, not the Foucauldian archive as a historical 'mass of said things' comprising multiple medialities.

In an article published in *Le foucauldian*, a journal for 'Research along Foucauldian Lines', Maurice Erb, Simon Ganahl, and Patrick Kilian explore how Moretti's distant reading and Foucauldian discourse analysis can be combined. To their main question of whether 'historical discourse analysis as practiced by Michel Foucault [can] be carried out with the aid of computers' (2016, p. 2)—the authors conclude that it requires that we become able to operationalize Foucault's concepts, particularly the concept of the statement. Yet they specify some difficulties in this regard. First, according to them, Foucault does not provide a clear answer to the question of what a statement is in *L'archéologie du savoir* (2016, p. 6). Second, Foucault's concept of the statement 'seems deliberately to block its operationalisation through digital analysis procedures'[2] because 'the archeological method [. . .] aims at a structural analysis not of the signifier but of the signified, thus that which is referred to and not the chains of signs with which both structuralism and text mining are concerned' (2016, p. 6). Erb *et al.* are right in pointing out that the Foucauldian statement has nothing to do with statistical signifiers, whereas text mining obviously does. They have in this regard highlighted important challenges. Contrary to their claim, however, Foucault does in fact specify what a statement is, as I will show in this section. Furthermore, as I argue in the next section, the Foucauldian statement can be operationalized in relation to digital research methods; text mining can be used to capture enunciations, from which we can identify statements.

In his description of the statement, Foucault demonstrates that it is distinct from linguistic elements such as the proposition, phrase, and speech act. The statement '[i]n its way of being unique (neither entirely linguistic, nor exclusively material) [. . .] is indispensable in order to say whether or not there is a sentence, proposition, or speech act; and whether the sentence is correct (or acceptable, or interpretable), whether the proposition is legitimate and well constructed, whether the speech act fulfils its requirements, and was in fact carried out' (Foucault 1972, 86).[3] Statements are thus something beyond these linguistic elements. Statements for Foucault enable us to articulate their existence and evaluate them (Foucault, 1969; Eliassen, 2016; Karlsen, 2020). There is a singular way that the statement makes certain ensembles of signs exist; it does so through the specific relation it has to 'something else': 'a series of signs will become a statement on condition that it possesses "something else" [. . .], a specific relation that concerns itself—and not its cause, or its elements' (Foucault, 1972, p. 89). The archive is precisely an epistemological object within which the statement and its relation to 'something else' can be identified.

The specific function a statement has, is dependent on its relation to other statements within the archive. The notion that a statement is context-dependent can for instance be seen in one of Foucault's examples in which the proposition 'dreams realize desires' is a different statement with a different function in Freudian psychoanalytical discourse than it is in platonic discourse.

An archival analysis will show that a limited number of things can be said, in limited ways and contexts, and what status is required status in order to access discourse. Foucault states: 'To define a system of formation in its specific individuality is [. . .] to

characterise a discourse or a group of statements by the regularity of a practice' (1972, p. 74). The regularity in ways of speaking informs the laws of possibility and rules of existence of a statement. When identifying the regularity in terms of what is said, how, and by whom[4] we can study what functions these ways of speaking have. The different functions of the statement can be specified through the identification of other statements by which this statement is surrounded in the respective archive.

To sum up, the statement in the Foucauldian sense delimits what can be said in a given historical period and conditions whether we experience a given sentence as correct or acceptable. Furthermore, the statement is context-dependent—it has a specific function due to its relation to other statements identified in the given material. Finally, the statement is an epistemological object that is identified by observing patterns in terms of who speaks, what is said, and how. In order to identify such regularities of ways of speaking, Foucault's distinction between the statement and the enunciation is crucial.

4 Statement and Enunciation

In the Foucauldian method, enunciations are unique realizations of a statement, whereas statements delimit what can be enunciated in a given historical period. The multiplicity of enunciations, for instance a person's repetitions of the same phrase, does not correspond to the same number of statements. The same statement can give rise to several enunciations; each enunciation is a concrete event that cannot be repeated (Foucault, 2014/1972, pp. 140/102). Thus, even if the statement also has a material existence in Foucault's definition of it—meaning that the statement does not exist independent of the time, place and materiality it appears in—it can be repeated, and it exists on a different level from the enunciation. Enunciations are given, the author of the formulation is the same as the subject of the enunciation, whereas the statement is the identification made by the researcher of the regularity across these enunciations (Karlsen, 2020).

Although the statement is neither necessarily the same at two different moments where the same linguistic signs come into existence, nor defined by one specific combination of words and signs and only these, the constancy of the statement within an archive differs from that of the enunciation: '[T]he constancy of the statement, the preservation of its identity through the unique events of the enunciations, its duplications through the identity of the forms, constitute the function of the field of use in which it is placed' (1972, p. 105). As this 'constancy' of the statement does not correspond to statistical signifiers, they cannot be directly identified in a collection of documents through the use of data mining techniques that register word frequency and the simple co-occurrence patterns of words. One can, however, use digital research methods to capture a concentrated mass of text passages from a digital collection, for instance passages that speak about sexuality or welfare (within a given period/area relevant to the researcher's focus). Such a concentrated material will be the researcher's archive.

Although enunciations, as concrete realizations of statements, will vary in their verbal combination, I will argue in the following that they can be captured through the use of sub-corpus topic modelling along with a bag of word tool. One can study these passages as a 'mass of said things' or enunciations, from which one can identify statements. Through examining these enunciations, we can identify statements which govern what it is possible and/or permissible to say under the specific historical conditions in question. In other words, one can identify something constant that runs through a certain multiplicity of enunciations that are captured in a material and that name, designate or describe a certain subject. In the example I will provide, I have delimited multiplicities of such enunciations—on women that I have harvested from the Norwegian National library's digitized book collection—and studied it as the field of use of the statements that I have identified.

5 Methodology

The overarching criterion for a relevant digital methodological design for the identification of discourse analysis in Foucault's tradition is that it must be able to generate a concentrated group of passages that articulate the subject matter in question. Let us imagine that a researcher has a hypothesis that a certain period of more or less thirty years in a determined area should be analysed with regard to the emergence

of a social practice, intellectual movement, set of social norms/customs/laws, and/or form of scientific knowledge. The larger and more diverse the amount of documents made available to this researcher is, the clearer the advantage to her research output. In principle such an analysis can be conducted using digital methods if the researcher has access to a rich digital collection of a great variety of documents, such as different forms of non-fictional and fictional literature, prestigious, canonized, and popular literature as well as less known, ignored or forgotten texts, for instance newspapers, pamphlets, scientific documents, magazines, journals, political documents, religious literature, legal documents. In order to identify discourse across such a digital collection of documents, the main challenge is to capture paragraphs that articulate the subject matter of interest—through different wording—in as complete and concentrated a manner as possible (free from 'noise', i.e. irrelevant passages). This concentrated material will be the Foucauldian archive as outlined above. To achieve this material, a combination of sub-corpus topic modelling (STM) and a bag of word tool is a valid approach. (This section details this approach and Fig. 1 summarizes its different steps.)

5.1 STM—Generation and selection of topics

In their pioneer article 'Trawling in the Sea of the Great Unread: Sub-corpus topic modelling and Humanities research', Peter Leonard and Timothy R. Tangherlini state that sub-corpus topic modelling (STM) 'increases the ability to discuss aspects of influence and intellectual movements' (2013, p. 725). In what follows, I will build on their work to show the suitability of STM for identifying discourse in the Foucauldian tradition. Briefly put, the first step for the researcher is to select one or several documents that she considers to be relevant examples from the period that articulate the research subject in question. This or these documents will serve as the sub-corpus/sub-corpora.

After several forms of curation, topics are generated, for example using an implementation of LDA (Latent Dietrich Allocation).[5] Topics are probability distributions of the words in the documents in the sub-corpus. Relevant topics for the research question are applied to the target corpus (the 'Great Unread' or large digital collection of 'all' documents from the period of interest). Automatically generated topics are probability distributions of the words—or 'discourses'—that could have generated the documents. We might have selected a document as sub-corpus that does not only address the subject that interests us. If we add all the topics to the target corpus, multiple captured results may be irrelevant. The aim, however is to capture a concentrated material of passages in the target corpus that articulate more or less the same subject matter as the passages in the sub-corpus. Thus all the generated topics from a given sub-corpus will not necessary be relevant to apply to the target corpus. As always with topic modelling, meticulous experimentation is necessary regarding the number of topics that are generated and what topics (when conducting sub-corpus topic modelling) that are applied to the target corpus.

It is, however, important to have in mind that Foucauldian discourse analysis deals with ensembles of interrelated statements and that their functions are determined by these relations. When examining the generated topics, the researcher should have in mind that topics comprising surprising combinations of words might in some cases indicate unexpected relations between subject matters. Not discarding topics too quickly might lead the researcher to identify unexpected statements and their co-relations to other statements. One might for instance identify that ways of speaking about health have a certain function in an archive by its relation to ways of speaking about another, supposedly unrelated, subject. Through the examination of these ways of speaking one may discover that something can be said given that it for example is moderated or contextualized in specific ways, whereas other alternatives are excluded.[6]

It is not sufficient, however, to select apt documents to use as sub-corpora, have topics generated from them, and apply these topics to the target corpus. First of all, topic generation and sound application of these topics to the target corpus demand careful curation and experimentation with different parameters. Leonard and Tangherlini specify six steps in the STM process: selection of sub-corpus, chunking, topic generation, curation, expert input and output, application of topics to target corpus (2013, p. 730). Furthermore, when our aim is to identify discourse, capturing

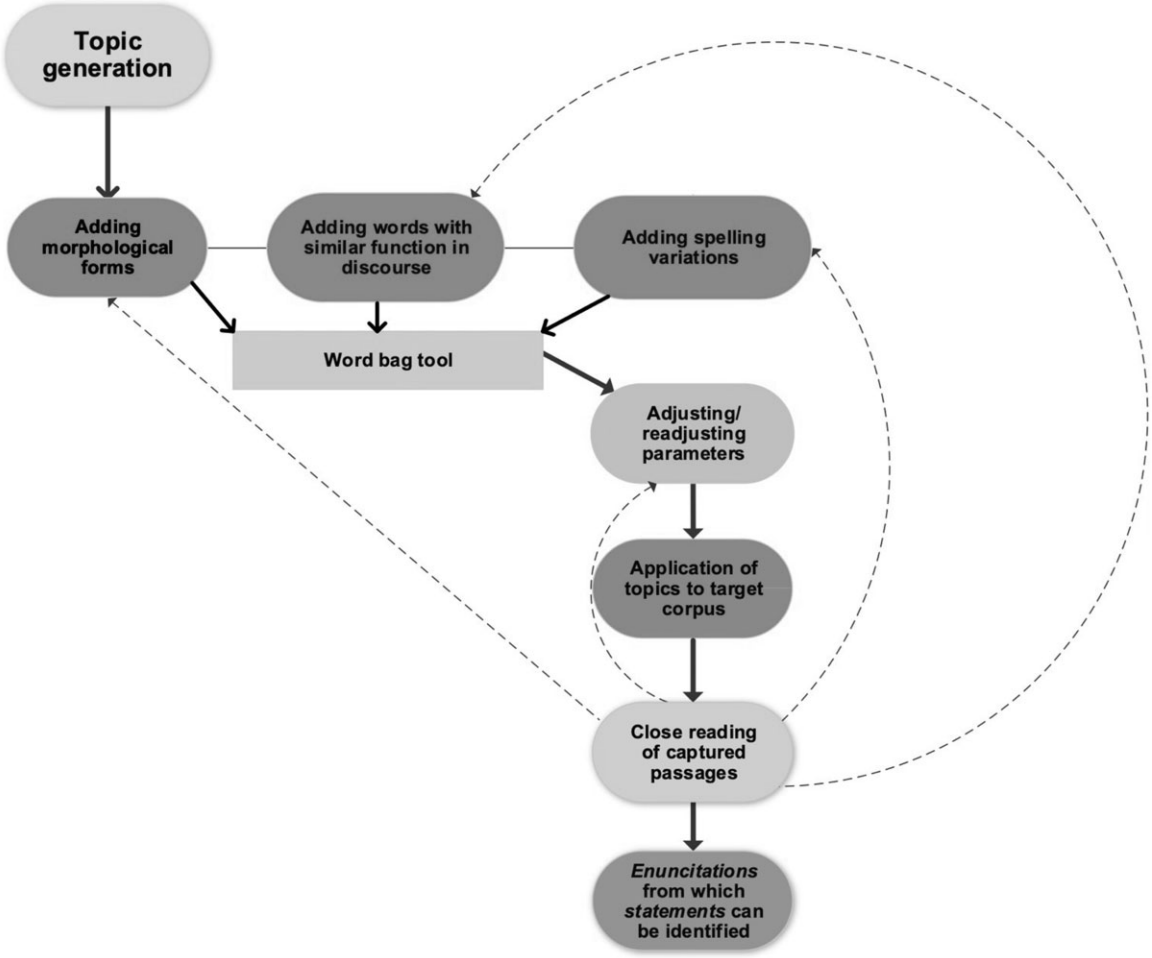

Fig. 1 The steps in the STM and bag of word tool process for obtaining a concentrated material of enunciations, based on which discursive statements can be identified.

passages that contain the same 'semantic feel', which Leonard and Tangherlini demonstrate that STM can do, is not sufficient. The topics generated by the algorithm contain specific combinations of linguistic signs (words in general). We have seen, however, that a Foucauldian statement does not necessarily come across through one and only one combination of linguistic signs. Despite the possible use of lemmatizers in order to make the topic model sensitive to words with the same root as the topic words—for instances different tenses of a verb—the passages that the topic model will capture are limited to the word register used in the document in the sub-corpus in question. It will, however, be indispensable to capture passages that articulate more or less the same subject matter and have more or less the same function in discourse, despite different wording. Not only synonyms to the topic words will be necessary to take into account, but words that differ in semantic content as well. We can for instance imagine that the words 'immigrant' and

'refugee', despite their different meanings, might have more or less the same function in a certain discourse.

5.2 Bag of word

A bag of word tool is a good solution to the challenge of preparing a topic model that captures passages that share the same function in discourse as the documents in the sub-corpus, although containing a different word register. Based on the researcher's domain expertise, he can add words that he has experienced or hypothesizes have a similar function in the discourse he is about to identify as the topic words in question. More specifically, the researcher adds relevant words for each topic word, in order for the topic model to handle these added words as if they were occurrences of the topic words in question when the algorithm goes through the documents in the target corpus. The words that are added might be spelling variations of the topic words (for instance due to anachronistic word forms or if one works with material with optical character recognition (OCR) problems), or—and this is very important when the aim is to prepare a material for Foucauldian discourse analysis—words with a similar function in discourse as the topic word in question.

One might of course object that one does not know how different articulations function in the discourse that one sets out to identify. A way around this challenge is the 'toggling' between distant reading and close reading. It is important to emphasize that digital methodology has to be combined with close reading of the material it captures throughout the process. One applies an initial set of topics—with correspondent bags of words for each topic word where it is relevant—and analyses the results they produce. The researcher has to pay attention to concrete word use in the captured passages, as well as to patterns in terms of who speaks and how, and to what these ways of speaking produce. The latter could for instance be that the human body becomes an object of knowledge and power, as Foucault observed (1976).

The researcher thus has to evaluate the results in accordance with specific criteria, but also in light of her domain expertise. The observations the researcher makes of the concrete words conveying the articulations in question, serve to inform her as to words that should be added to the word bags. Thus, if the researcher observes articulations in the target corpus with more or less the same discursive function as the topic word of her sub-corpus, she will need to refine her topic model by adding this word to the relevant bag of words. After this refinement of the topic model, new STM searches need to be conducted.

5.3 STM—application of topics to target corpus

The application of the topics to the target corpus (the chunking and the parameterization) has to be further specified. As we have seen above, the objective is to capture what we in Foucauldian terminology call enunciations. From these enunciations the overarching task is to identify the statements of which the enunciations are concrete realizations. We saw that Foucault uses the example 'dreams realise desires' to highlight that the same combination of linguistic signs can correspond to entirely different statements, belonging to different discourses. Statements are context-dependent, or more precisely, they always have a relation to 'something else'. Thus, we cannot, for instance, program the topic model to chunk the documents into phrases. Even if there are phrases in the target corpus in question that comprise all the topic words of the topic in question (or, more precisely, words from each bag of words corresponding to each topic word of the topic in question), such results will not be sufficient to identify statements in the material. Instead, it is much more suitable to chunk the target corpus into paragraphs. In many cases a paragraph provides enough context to determine what is being articulated, by whom (i.e. who has access to discourse/what is required in order to produce discourse), and how. Furthermore, when a paragraph is analysed in light of other captured paragraphs, the researcher is likely to be able to determine the pattern in terms of the functions these ways of speaking have.

Another important issue to settle is the degree of topic match we require in order for a paragraph in the target corpus to qualify as a hit and thus be captured by the topic model. What do we program the topic model to 'require' in order for a passage in the target corpus to qualify as a hit? When the topics from the sub-corpus are applied to the target corpus, several procedures are possible.

One option is to take the Jaccard index into account.[7] This entails that the occurrences of the topic

words are counted in every passage in the documents in the target corpus and the passages are then compared to one another. Based on this counting and comparison, a Jaccard score is given to each passage (Karlsen, 2020). One can experiment with different Jaccard settings. When identifying a discourse, one needs to stay alert to other possible word combinations and to encode for other n-grams than the ones one was initially able to predict. In order to search broadly for potentially relevant passages, an appropriate strategy is to lower the Jaccard similarity value. One can for instance program the topic model to register passages as hits if at least one word is identified from four out of a total of six word bags (that each correspond to a topic word). One can also use a 'conditional Jaccard', where some specific topic words (or words form their correspondent word bags) must be identified in order for the passage in question to be captured, in addition to for example three out of any of the remaining six.

The challenge the researcher has to address, through trial and error and based on her domain expertise, is to find the settings that avoid a lot of irrelevant results, but at the same time capture relevant passages, including passages with a word register that one had not been able to anticipate. In some cases, the researcher might also discover that the time period designated for the archive should be slightly modified.

The overarching objective of this trial—and—error process with different parameters and adjustments of the Jaccard setting is to obtain a concentrated material of passages—from well-known texts to unknown texts from the 'Great Unread'—that are relevant for the discourse one sets out to identify. These passages can be considered enunciations in the Foucauldian sense that I have specified above.

5.4 From enunciations to statements

The researcher's next task is to analyse the patterns in these ways of speaking. This includes regularities in terms of who speaks and how, and what these ways of speaking produce, as well as the relations and co-dependence of articulations of norms, customs, knowledge, etc., and thus identify the statements of which the mass of captured enunciations can be considered concrete actualizations.

In addition to close reading of the results, one might consider using structured topic modelling to train a topic model in order to approach some of these questions. One can add confounding factors to the topic model, for instance for the analysis of the attitude or sentiment encoded in passages, authors, variations in topics across the texts from which the passages have been captured, as well as insight browsers that allow the researcher to study the prevalence of topics across the entire corpus.[8] In the examination of a set of captured material, word frequency statistics, concordance and collocation analyses are also useful in order to map regularities in ways of speaking.[9] Such analyses must, however, be complemented by the researcher's close reading of the material. The last part of this article presents how I have used the methodology described above to identify a discourse on women in books from 1830 to 1880 digitized by the Norwegian National Library.

6 Women's 'place' in Norwegian Society 1830–80

The emergence of What Foucault has labeled population politics has been much discussed in political philosophy and history. Population politics entails that the population is a political subject. It is a level for economic and political action (Foucault, 2004, pp. 32–50). Population politics applies to Norway as to many other western countries in the period. A relevant example is that women's housework gained more attention (and esteem in a certain sense) because the authorities and enlightened elite were aware of its importance for the population's health (Hagemann, 2005, p. 191; Karlsen, 2020). Concurrent with the growth of population politics in the nineteenth century, reforms regarding political rights for men take place in numerous western countries. A public sphere for negotiating political rights for men emerges, whereas women are almost entirely absent from this sphere. Although women work in the industry, as telegraphists and teachers, etc., it is well established in previous research that women's 'place' is considered to be primarily in the private home.

The discourse I have identified, however, enables a space for negotiating and eventually resisting women's place in society. This establishment of women's place is part of a discourse related to population politics and

to a Christian, bourgeois feminine ideal. Although this discourse establishes women's sphere as other to the public sphere where men's political rights are discussed, a function of the same discourse is paradoxically that a space leading up to the women's movement from the 1880s emerges. My application of the Foucauldian digital discourse analysis yielded more historically revealing results than one might otherwise find. Without such an approach, that integrates a large number of documents of different genres and from different disciplines, we might be blinded to some of the longstanding and piecemeal changes that actually led to the women's movement. An example is a large number of religious texts, by Norwegians or translated. Analysed together, these texts testify to a double function: a moralization of woman and her place in the private sphere on the one hand, and an invitation for women to engage in the discourse of her place on the other hand.

The Norwegian National library has digitized more than 500,000 books and 3 million newspapers. Through an API created for researchers, a large part of the material can be text mined.[10] The book corpus in this digital collection comprises publications dating from the eighteenth century up until today. Taking advantage of the access to this 'Great Unread' of texts published in Norway, I have studied ways of speaking about women and how women themselves increasingly spoke about their gender between 1830 and 1880. As in many Western countries, what is often referred to as the first wave women's movement took place in Norway in late nineteenth and early twentieth century. In Norwegian history, the 1880s is a period marked by an accelerating women's emancipation movement. These years have been subject to extensive interdisciplinary research.[11] The conditions of women and articulations of woman's 'place' in society in the decades preceding this period have not, however, been analysed extensively. Briefly put, concurrently with the emergence of the public sphere and political reforms (for men) in nineteenth-century Norway and many other Western countries, woman's 'place' in society was increasingly specified as the private home (Hagemann, 2005; Karlsen, 2020). With the aim of shedding light on ways of speaking about women, articulations and negotiations of woman's 'place' in society, and the increase of women writers who thematized their own gender starting in the 1830s, I conducted an analysis of the discursive establishment of woman's 'place' in society in the period 1830–80.[12]

Sub-corpus topic modelling permitted me to examine topics from selected texts in order to search for occurrences of the same topics—i.e., passages that speak about woman and gender—in the Norwegian National Library's digital collection. I started with novels and newspaper articles by selected women writers and modelled topics from these texts.[13] I used NMF (Non Negative Matrix factorization) and LDA (Latent Dirichlet Allocation) as topic modelling algorithms and created topics manually as well (see below). As part of curating the topic model, I added morphological forms of the topic words. Tools for automatic lexemization could have been be used. There are, however, so many spelling variants of the topic words (due to OCR errors) that have to be accounted for that such tools would not have been very helpful. Furthermore, the texts in the target corpus are only tokenized down to word level (in 2021). Other types of pre-processing such as lemmatization and lowercasing have not taken place. Every potential, unique combination of characters in the target corpus that interests the researcher thus has to be specified in the topic model.

The initial applications of the topics to the target corpus produced few results. I experimented with lowering the Jaccard similarity value, which means that I had the parameters reset so that a lower percentage of the topic (number of topic words or words from their correspondent word bag) had to be identified in a passage in the target corpus in order to qualify as a hit. From the results this readjusted topic model produced, I realized that the OCR had produced multiple spelling variations of words among the topic words.[14] As these had not been taken into account in the topic curation, they were not registered by the topic model.

Furthermore, many of these preliminary results provided information about different words that more or less conveyed the same meaning, or were used the same way as words among the topic words, for instance multiple words signifying woman that I had not thought of. In sum, although some of these results were irrelevant because the Jaccard similarity value was set low enough to include passages that did not concern women, the results were important in order to become aware of words or spelling variations

of words that should be added to the word bags, followed by new searches.

In order to obtain a rich and concentrated collection of paragraphs, I ran many searches, continually readjusting the topics and their correspondent word bags to fit my examinations of the captured passages. Using STM in order to capture material for conducting discourse analysis is not a manner of investigating whether the texts in the sub-corpora have had an impact on other texts, or if the 'semantic feel' in the texts in the sub-corpora can be found in texts in the target corpus. Nevertheless, in the creation of the 'archive' in this example, it was essential to start off from certain texts from the period in order to map words that were used to speak about women. Yet it was equally essential to gradually move slightly away from them, in the sense of adapting the topics and word bags based on captured passages during the process.

It is possible to create topics manually as well, based on close reading of the texts in question, secondary literature on the period, etc. It is not essential that all topics be automatically created. One advantage of automatic topic generation that Leonard and Tangherlini highlight, however, is that one can capture passages comprised of relevant words that one had not initially associated with one's research question through preliminary close reading of the text in question (2013, p. 728). I agree with this position. The algorithm registers how the words are actually used in a given text in terms of which words tend to occur at the same time (Blevins, 2010). Although the researcher may know the texts in the sub-corpora well, the algorithm may calculate word co-occurrence patterns of which the researcher was previously unaware. Automatic topic generation is for this reason a valuable tool when applying Foucauldian discourse analysis. Used well, it facilitates the capture of text in the target corpus beyond our preconceived understanding of ways of speaking in the period. Nevertheless, the most important factor is to ensure that one captures the 'mass of said things' related to one's research topic. Manually created topics, based on the researcher's domain expertise and careful observations of the captured material during the working process, might be equally valuable and sometimes essential to add.

In the example project, I carried out a 'distant reading' of approximately 7,000 books (the number of books in the Norwegian National library's collection between 1830 and 1880). My final material, or 'archive', after the trial—and—error process comprised the set of paragraphs that I considered enunciations on women and gender. Through the observation of the regularity in ways of speaking across these paragraphs, I identified six interrelated gender discursive statements:

(1) Woman is potentially an exemplary Christian;
(2) Woman is of a sensitive nature;
(3) Woman must be persuaded to act in accordance with the virtues for her gender;
(4) Woman must be accorded freedom;
(5) Woman needs education in order to fulfil her calling; and
(6) Woman can improve the state of morality in her community.

It is beyond the scope of this article to elaborate on the variety of texts and genres in which the paragraphs were captured, the topics that captured them, the texts from which the topics were modelled, and the multiple ways these statements are related to one another. See Fig. 2 for the distribution of the main categories of genres/disciplines in which the passages were captured.

For multiple captured paragraphs, each one contributed to the identification of more than one discursive statement. The statement that woman must be accorded freedom, for instance, is a completely different statement, with a different function in this discourse, compared to for instance a second wave feminist discourse in the USA in the 1970s. Regarding the functions of these statements within discourse, note that there is a mixture of descriptive and normative statements. Through descriptions in the archive of woman's nature, a conception of natural gender difference is created, concurrently with articulations of how woman best can fill her presumed natural function in accordance with this feminine essence. Multiple works, didactic works in particular, address woman directly and function as attempts to persuade women to eagerly fulfil this 'natural' calling of their own accord. The main functions of the articulated importance of woman's attributed tasks were a cementation of woman's 'place' in the private home, but also that woman's freedom, virtues and education became open for negotiation.

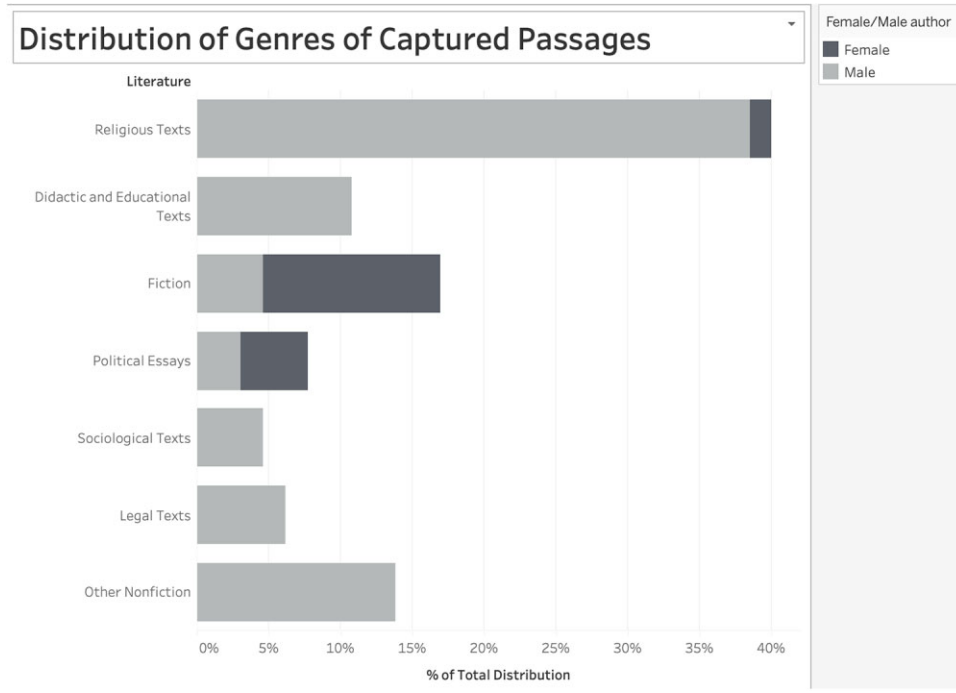

Fig. 2 Distribution of genres of captured passages. We see that religious texts constitute by far the most widespread category. Passages captured in didactic and educational texts have been important for the identification of the normative statement 'woman must be persuaded to act in accordance with the virtues for her gender'. This chart is based on a sample of the captured passages and show the main trends. There are texts we can qualify as didactic, written by women in the period as well. As the graph shows, however, it is first and foremost in fiction and political essays that passages by women authors have been captured.

The archival context of these ways of speaking about woman in the example case is dominated by population politics.[15] These ways of speaking emerge with growing knowledge about the state of the population brought about by new technologies and new disciplines, such as sociology and demography, as well as with improvement in education, advocacy for rights, new means to reach more people with information, and concerns regarding the stability of institutions. Woman has a function in a paradoxical landscape where freedom has to be reconciled with specific virtues. Women were thought to have had a predetermined essence that ought nevertheless to be pursued freely. The women's movement of the 1880s was not a phenomenon that could come about through a sudden rupture with the Christian and bourgeois feminine ideal. Indeed, what has often been described as a sudden rupture was nothing of the sort. Early ideas that would eventually culminate in the liberation movement of the 1880s started brewing within conservative bourgeois circles long before. The discursive establishment of women's place in society in the period 1830–80, all the while cementing this 'place' as a gender-specific function tied primarily to the private home, comprised negotiations of women's freedom and education. One function of that was the creation of a space for and empowerment of woman to engage in the negotiation of her 'place'—and eventually resist it (Karlsen, 2020). In spite of taking place within seemingly misogynist terms (from a contemporary perspective), the Norwegian discourse on women from before the 1880s thus gradually allowed women to engage in the negotiation of their 'place', which became a precondition for their eventual ability to resist.

Digital methods alone could not have fully detected these findings. Qualitative analyses of the ways of speaking in the captured passages, as well as in terms of the genre, authorship, etc., of the texts from which they are captured, are necessary as well to deduce the functions of the statements within the analysed archive.

7 Conclusion

Large digitized collections of texts and digital research methods offer new opportunities for conducting discourse analysis in the Foucauldian tradition. Although the Foucauldian archeological method has statements and not language—or statistical signifiers—as its object of study, a combination of sub-corpus topic modelling and a word bag tool can be used to gather a set of enunciations (énonciations) from which one can identify discursive statements (énoncés). Starting from well-known texts, using topics modelled from them to search a large, less known target corpus, the procedure consists of gradually improving the topic model by revising the word bags as one step by step increases one's knowledge of the 'mass of said things' through analyses of the captured material along the process. This process includes mapping the multiplicity of words that are used in the different ways of speaking about the subject matter of the analysis. This mapping serves to refine the topic model, re-apply the topic model to the target corpus, and finally to analyse the mass of captured paragraphs, which can be regarded as enunciations. Through identifying regularities in these ways of speaking, one can deduce the various functions they have within a given archive. With perseverance, we may reach significant new historical insights using this methodology.

Acknowledgements

The help I have got from computational linguist Lars G. Johnsen, who also works as a programmer at the Norwegian National Library, has been an indispensable. I would also like to thank Timothy R. Tangherlini, Ellen R. Rees, and Else C.W. Werring for insightful comments.

References

Blevins, C. (2010). Topic modeling Martha Ballard's diary, *Historying*. http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary/ (accessed 1 December 2019)

Eliassen, K.O. (2016). *Foucaults begreper*. Oslo: Sap.

Erb, M., Ganahl S., and Kilian, P. (2016). Distant reading and discourse analysis. *Le foucaldien*, **2**/1, DOI: 10.16995/lefou.16.

Foucault, M. (2014) [1969]. *L'archéologie du savoir, Tel*. Paris: Éditions Gallimard.

Foucault, M. (1972). *The Archeology of Knowledge*. New York: Pantheon Books.

Foucault, M. (1976). *Histoire de la sexualité: 1: La volonté de savoir*. Vol. **1**. Paris: Gallimard.

Foucault, M. (2004). Sécurité, territoire, population: cours au Collège de France, 1977-1978. In Senellart, M. et al. (eds), *Hautes études*. Paris: Seuil Gallimard.

Hagemann, G. (2005). De stummes leir? 1800-1900. In Blom, I. and Sogner, S (eds), *Med kjønnsperspektiv på norsk historie. Fra vikingtid til 2000-årsskiftet*. Oslo: Cappelen Akademisk forlag, pp. 157–255.

Iversen, I. (1988). Et moderne gjennombrudd. In Engelstad, I. et al. (eds), *Norsk kvinnelitteraturhistorie Bind 1 1600-1900*. Oslo: Pax, pp. 155–68.

Jo, W. (2019). Possibility of discourse analysis using topic modeling. *Journal of Asian Sociology*, **48**(3): 321–42.

Karlsen, H. (2020). *A Discourse Analysis of Woman's Place in Society 1830–1880 through Data Mining the Digital Bookshelf*. Ph.D. Dissertation, University of Oslo.

Sam, C. H. (2019). Shaping discourse through social media: using Foucauldian discourse analysis to explore the narratives that influence educational policy. *American Behavioral Scientist*, **63**(3): 333–50. DOI: i1.o0r.g1/107.171/0770/020072674624128188220056 5

Tangherlini, T. R. and Leonard Peter. (2013). Trawling in the sea of the great unread: sub-corpus topic modeling and humanities research. *Poetics*, **41**(6): 725–49.

Winthrop-Young G. (2015). Discourse, media, cultural techniques: the complexity of Kittler. *MLN*, **130**(3): 447–65.

Notes

1 My translation of the original: 'Le domaine de choses dites, c'est ce qu'on appelle l'archive; l'archéologie est destinée à en faire l'analyse' (*L'archéologie du savoir*. 2014 [1969]. Paris: Gallimard edition Tel).

2 Here Erb, Ganahl, and Kilian cite and translate from German Peer Trilcke and Frank Fischer. 'Fernlesen mit Foucault, Überlegungen zur Praxis des distant reading und zur Operationalisierung von Foucaults Diskursanalyse'. In *Le foucauldian*, 2016. 2/1, p. 17.

3 'Il est, dans son mode d'être singulier (ni tout à fait linguistique, ni exclusivement matériel), indispensable pour qu'on puisse dire s'il y a ou non une phrase, proposition, acte de langage ; et pour qu'on puisse dire si la phrase est correcte (ou acceptable, ou interprétable), si la proposition est légitime et bien formée, si l'acte est conforme aux réquisits et s'il a été bel et bien effectué' (Foucault, 2014, p. 119).

4 Identifying discourse does not include interpretation of concrete enunciations based on the author's biography or other elements that would individualize what is said. Instead, the relevance of the author concerns the status an individual must have in order to access discourse and become the subject of enunciation of the discursive statement in question.

5 See for instance Leonard and Tangherlini (2013). BERT models, TF-IDF and Non-Negative Matrix Factorization are other examples. It might be useful to experiment with several models on the same material.

6 See for instance Geoffrey Winthrop-Young on how Friedrich Kittler explored this complexity in Foucauldian discourse analysis (Winthrop-Young, 2015) .

7 The Jaccard index, or Jaccard similarity coefficient, is a concept in statistics, referring to the measure of a compared similarity of a certain amount of samples, named after the Swiss statistician Paul Jaccard.

8 See Molly Roberts, Brandon Stewart, and Dustin Tingley https://github.com/bstewart/stm

9 This can be done with a tool for text mining such as Voyant Tools. See Stéfan Sinclair & Geoffrey Rockwell: https://github.com/sgsinclair/Voyant.

10 The Norwegian National Library's DH-Lab: https://github.com/DH-LAB-NB/DHLAB

11 See for instance Hagemann (2005) and Iversen (1988).

12 See my PhD dissertation (Karlsen, 2020) for a complete presentation of this project and its documentation here: https://github.com/heidikarlsen/Documentation_Dissertation

13 I did not limit myself to Norwegian women writers. This is because I started from the assumption that the discourse on women, preceding the first wave women's movement, is international. I have conducted studies of the French writer George Sand's novel *Indiana* (1832), the Swedish writer Fredrika Bremer's newspaper article 'To the women of Sweden' (1844) and her novel *Hertha* (1856), the Norwegian writer Camilla Collett's novel, *The District Governor's Daughters* (1854/55), and the Norwegian Aasta Hansteen's series of newspaper articles ' Women's opinion on "women's subjection"' (1870). *The District Governor's Daughters* was published in 1854 and 1855, thus in the very middle of the period 1830–80. Collett's novel is considered the first Norwegian modern novel. The main theme in the novel is the question of woman's free will versus the persuasion to which she is subjected (Karlsen, 2020). Collet refers to Bremer and Sand and took inspiration from them. (See also my article on Sand and her importance for the women's movement in Norway: 'Le mouvement des femmes en Norvège au XIXe siècle avait-il besoin de George Sand pour sortir de son silence ?' 2019. In *Deshima* (13), pp 195–211.). The works that I have selected to study in this project serve as context for Collett, so to speak. Hansteen was later to become a well-known activist in the women's movement. I study her polemic series of articles as exemplifications of women's access to gender discourse though writing in newspapers. Her articles, in which she defends John Stuart Mill's *The Subjection of Women* (1869), contribute to an important debate that mobilised several women to write articles in the press.

    Example of applied topics with translated topic words (based on which I have created bag of words) in parenthesis are: 'woman's function in the public sphere' (woman, fatherland, chastity, deprivation), 'Is love true?' (love, true, heart, eye, self, woman), 'Woman's being and acquired traits' (woman, character, mores).

14 Working with digitized material from the nineteenth century (or before) often entails a significant problem, namely OCR errors. Numerous documents were printed in fraktur (a variant of Gothic typeface), which might lead to poor character recognition in the digitization process. In order to reduce the risk that potentially relevant passages went unnoticed because of spelling variations due to OCR errors, I ran word searches with wildcards in the target corpus. This wildcard search algorithm is based on indexing

and inspired by the Levenshtein distance metric. Wild card search in the target corpus can be used for improving the word bags, so that actual misspellings of the topic words in the corpus are constitutive for the word forms one adds to the word bags. One can include several wildcards in the search word. Letters that are vulnerable to misreading are replaced by wildcards. In some cases one letter is misread as two characters, etc., which can be taken into account by programming the algorithm to search for all compatible words in the database where the Levenshtein distance is up to two characters between the search word.

15 See Foucault (2004) on population politics in *Sécurité, territoire, population : cours au Collège de France, 1977-1978*, Hautes études, edited by Michel Senellart *et al.* Paris: Seuil Gallimard.