Abstract for long format presentation
DHN2018

Heidi Karlsen, University of Oslo
Ph.D. Candidate in literature,
Cand.philol. in philosophy

Interdisciplinary advancement through the unexpected: Mapping gender discourses in Norway (1840-1913) with *Bokhylla*

This presentation discusses challenges related to sub-corpus topic modeling in the study of gender discourses in Norway from 1840 till 1913 and the role of interdisciplinary collaboration in this process. Through collaboration with the Norwegian National Library, data-mining techniques are used in order to retrieve data from the digital source, *Bokhylla* [«the Digital Bookshelf»], for the analysis of women's «place» in society and the impact of women writers on this discourse. My project is part of the research project «Data-mining the Digital Bookshelf», based at the University of Oslo.

1913, the closing year of the period I study, is the year of women's suffrage in Norway. I study the impact women writers had on the debate in Norway regarding women's «place» in society, during the approximately 60 years before women were granted the right to vote. A central hypothesis for my research is that women writers in the period had an underestimated impact on gender discourses, especially in defining and loading key words with meaning (drawing on mainly Norman Fairclough's theoretical framework for discourse analysis). In this presentation, I examine a selection of Swedish writer Fredrika Bremer's texts, and their impact on gender discourses in Norway.

The Norwegian National Library's Digital Bookshelf*,* is the main source for the historical documents I use in this project*.* The Digital Bookshelf includes a vast amount of text published in Norway over several centuries, text of a great variety of genres, and thus offers unique access to our cultural heritage. Sub-corpus topic modeling (STM) is the main tool that has been used to process the Digital Bookshelf texts for this analysis. A selection of Bremer's work has been assembled into a sub-corpus. Topics have then been generated from this corpus and then applied to the full Digital Bookshelf corpus. During the process, the collaboration with the National Library has been essential in order to overcome technical challenges. I will reflect upon this collaboration in my presentation. As the data are retrieved, then analyzed by me as a humanities scholar, and weaknesses in the data are detected, the programmer, at the National Library assisting us on the project, presents, modifies and develops tools in order to meet our challenges. These tools might in turn represent additional possibilities beyond what they were proposed for. New ideas in my research design may emerge as a result. Concurrently, the algorithms created at such a stage in the process, might successively be useful for scholars in completely different research projects. I will mention a few examples of such mutually productive collaborations, and briefly reflect upon how these issues are related to questions regarding open science.

In this STM process, several challenges have emerged along the way, mostly related to OCR errors. Some illustrative examples of passages with such errors will be presented for the purpose of discussing the measures undertaken to face the problems they give rise to, but also for demonstrating the unexpected progress stemming from these «defective» data. The topics used as a «trawl line»[1], in the initial phase of this study, produced few results. Our first attempt to get more results was to revise down the required Jaccard similarity value[2]. This entails that the

[1] My description of the STM process, with the use of tropes such as «trawl line» is inspired by Peter Leonard and Timothy R. Tangherlini (2013): "Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and Humanities research" in Poetics. 41, 725-749

[2] The Jaccard index is taken into account in the ranging of the scores. The best hit passage for a topic, the one with highest score, will be the one with highest relative similarity to the other captured passages, in terms of concentration of topic words in the passage. The parameterized value of the required Jaccard similarity defines the score a passage must receive in order to be included in the list of captured passages from the «great unread».

quantity of a topic that had to be identified in a passage in order for it to qualify as a hit, is lowered. As this required topic quantity was lowered, a great number of results were obtained. The obvious weakness of these results, however, is that the rather low required topic match, or relatively low value of the required Jaccard similarity, does not allow us to affirm a connection between these passages and Bremer's text. Nevertheless, the results have still been useful, for two reasons. Some of the data have proven to be valuable sources for the mapping of gender discourses, although not indicating anything regarding women writers' impact on them. Moreover, these passages have served to illustrate many of the varieties of OCR errors that my topic words give rise to in text from the period I study (frequently in Gothic typeface). This discovery has then been used to improve the topics, which takes us to the next step in the process.

In certain documents one and the same word in the original text has, in the scanning of the document, given rise to up to three different examples of OCR errors[3]. This discovery indicates the risk of missing out on potentially relevant documents in the «great unread»[4]. If only the correct spelling of the words is included in the topics, potentially valuable documents with our topic words in them, bizarrely spelled because of errors in the scanning, might go unnoticed. In an attempt to meet this challenge I have manually added to the topic the different versions of the words that the OCR errors have given rise to (for instance for the word «kjærlighed» [love] «kjaerlighed», «kjcerlighed», «kjcrrlighed»). We cannot in that case, when we run the topic model, require a one hundred percent topic match, perhaps not even 2/3, as all these OCR errors of the same word are highly unlikely to take place in all potential matches[5]. Such extensions of the topics, condition in other words our parameterization of the algorithm: the required value of Jaccard similarity for a passage to be captured has to be revised fairly down. The inconvenience of this approach, however, is the possible high number of captured passages that are exaggeratedly (for our purpose) saturated with the semantic unit in question. Furthermore, if we add to this the different versions of a lexeme and its semantic relatives that in some cases are included in the topic, such as «kvinde», «kvinder», «kvindelig», kvindelighed» [woman, women, feminine, femininity], the topic in question might catch an even larger number of passages with a density of this specific semantic unity with its variations; this is an amount that is not proportional to the overall variety of the topic in question.

This takes us back to the question of what we program the "trawl line" to "require" in order for a passage in the target corpus to qualify as a hit, and as well to how the scores are ranged. How many of the words in the topic, and to what extent do several occurrences of *one* of the topic's words, i.e., five occurrences of "woman" in one paragraph interest us? The parameter can be set to range scores in function of the occurrences of the different words forming the topic, meaning that the score for a topic in a captured passage is proportional to the heterogeneity of the occurrences of the topic's words, not only the quantity. However, in some cases we might, as mentioned, have a topic comprising several forms of the same lexeme and its semantic relatives and, as described, several versions of the same word due to OCR errors. How can the topic model be programmed in order to take into account such occurrences in the search for matching

[3] Some related challenges were described by Kimmo Kettunen and Teemu Ruokolainen in their presentation, «Tagging Named Entities in 19th century Finnish Newspaper Material with a Variety of Tools» at DHN2017.

[4] Franco Moretti (2000) (drawing on Margareth Cohen) calls the enormous amount of works that exist in the world for «the great unread» (limited to *Bokhylla's* content in the context of my project) in: «Conjectures of World Literature» in New Left Review. 1, 54-68.

[5] As an alternative to include in the topic all detected spelling variations, due to OCR errors, of the topic words, we will experiment with taking into account the Levenshtein distance when programming the «trawl line». In that case it is not identity between a topic word and a word in a passage in the great unread that matters, but the distance between two words, the minimum number of single-character edits required to change one word into the other, for instance «kuinde»-> «kvinde».

passages? In order to meet this challenge, a «hyperlexeme sensitive» algorithm has been created[6]. This means that the topic model is parameterized to count the lexeme frequency in a passage. It will also range the scores in function of the occurrence of the hyperlexeme, and not treat occurrences of different forms of one lexeme equally to the ones of more semantically heterogenous word-units in the topic. Furthermore, and this is the point to be stressed, this algorithm is programmed to treat miss-spelling of words, due to OCR errors, as if they were different versions of the same hyperlexeme.

The adjustments of the value of the Jaccard similarity and the hyperlexeme parameterization are thus measures conducted in order to compensate for the mentioned inconveniences, and improve and refine the topic model. I will show examples that compare the before and after these parameters were used, in order to discuss how much closer we have got to be able to establish actual links between the sub-corpus, and passages the topics have captured in the target corpus. All the technical concepts will be defined and briefly explained as I get to them in the presentation. The genesis of these measures, tools and ideas at crucial moments in the process, taking place as a result of unexpected findings and interdisciplinary collaboration, will be elaborated on in my presentation, as well as the potential this might offer for new research.

[6] By the term «hyperlexeme» we understand a collection of graphemic occurences of a lexeme, including spelling errors and semantically related forms.