Abstract, long paper:

A Computational Assessment of Norwegian Literary "National Romanticism"

Ellen Rees, University of Oslo

In this paper, I present findings derived from a computational analysis of texts designated as "National Romantic" in Norwegian literary historiography. The term "National Romantic," which typically designates literary works from approximately 1840 to 1860 that are associated with national identity formation, first appeared decades later, in Henrik Jæger's *Illustreret norsk litteraturhistorie* from 1896. Gudleiv Bø has written extensively about numerous examples of national romanticism in Norwegian literature without probing the term itself to any great extent (Bø 1995, 1998 2006, 2008, 2011). Cultural historian Nina Witoszek has on a number of occasions written critically about the term, claiming that it is misleading because the works it denotes have little to do with larger international trends in Romanticism (see especially Witoszek 2011). Yet, with the exception of a 1985 study by Asbjørn Aarseth, it has never been interrogated systematically within the Norwegian context in the way that other period designations such as "Realism" or "Modernism" have.[1] Nor does Aarseth's investigation attempt to delimit a definitive National Romantic corpus or account for the remarkable disparity among the works that are typically associated with the term. "National Romanticism" is like pornography—we know it when we see it, but it is surprisingly difficult to delineate in a scientifically rigorous way.

Together with members of the project team, I have prepared a corpus of texts that are mentioned in connection with "National Romanticism" in the major histories of Norwegian literature in Norwegian literature. I will discuss briefly some of the logistical challenges associated with preparing this corpus.

This corpus forms the point of departure for a computational analysis employing various text-mining methods in order to determine to what degree the texts most commonly associated with "National Romanticism" share significant characteristics. In the popular imagination, the period is associated with folkloristic elements such as supernatural creatures (trolls, hulders), farming practices (shielings, herding), and folklife (music, rituals) as well as nature motifs (birch trees, mountains). We therefore employ topic modeling in order to map the frequency and distribution of such motifs across time and genre within the corpus. We anticipate that topic modeling will also reveal unexpected results beyond the motifs most often associated with National Romanticism. This process should prepare us to take the next step and, inspired by Matthew Wilkens' recent work generating "clusters" of varieties within twentieth-century U.S. fiction, create visualizations of similarities and differences among the texts in the National Romanticism corpus (Wilkens 2016).

Based on these initial computational methods, we hope to be able to answer some of the following literary historical questions:

---

[1] National romanticism in a central and eastern European context is treated in *National Romanticism: The Formation of National Movements* (Trencsenyi and Kopecek 2007).

- Are there identifiable textual elements shared by the texts in the National Romantic canon?
- What actually defines a National Romantic text as National Romantic?
- Do these texts cluster in a meaningful way chronologically?
- Is "National Romanticism" in fact meaningful as a period designation, or alternately as a stylistic designation?
- Are there other texts that share these textual elements that are not in the canon?
- If so, why? Do gender, class or ethnicity have anything to do with it?

To answer the last two questions, we need to use the "National Romanticism" corpus as a sub-corpus and "trawl-line" within the full corpus of nineteenth-century Norwegian textual culture, carrying out sub-corpus topic modeling (STM) in order to determine where similarities with texts from outside the period 1840–1860 arise (Tangherlini and Leonard 2013). For the sake of expediency, we use the National Library of Norway's Digital Bookshelf as our full corpus, though we are aware that there are significant subsets of Norwegian textual culture that are not yet included in this corpus. Despite certain limitations, the Digital Bookshelf is one of the most complete digital collections of a national textual culture currently available.

For the purposes of DHN 2018, this project might best be categorized as an exploration of cultural heritage, understood in two ways. On the one hand, the project is entirely based on the National Library of Norway's Digital Bookshelf platform, which, as an attempt to archive as much as possible of Norwegian textual culture in a digital and publicly accessible archive, is in itself a vehicle for preserving cultural heritage. On the other hand, the concept of "National Romanticism" is arguably the most widespread, but least critically examined means of linking cultural heritage in Norway to a specifically nationalist agenda.

Preliminary findings indicate that texts associated with "national romanticism" in the literary histories do in fact differ from the larger reference corpus, but not in the ways we expected. We anticipated clusters aligning with familiar national romantic themes (the supernatural, farming practices, folk life and nature motifs), and while these were indeed present, they were not the only markers for national romanticism. A quite different cluster emerged as equally, if not more, dominant, namely that of infatuation. It consists of words like: kiss, cheek, boy, girl, smile, beloved.

These preliminary findings are based on comparisons of word frequency in a target corpus consisting of 78 texts identified as national romantic and a randomly generated reference corpus of 500 books published in Norway between 1830 and 1890. The target corpus is much smaller than our original list because a number of works are shorter texts contained within the same book, such as folk tales or poems. To make the comparison, we generated wordlists for each of the two corpora and a third combined corpus. We then normalized word frequency so that word occurrence is relative to the book it appears in. The target and reference corpora were then aggregated so that they function as lists of word frequency. All words in each book were connected to an average relative frequency, so that the higher the number generated, the more specific the word is to that particular corpus. Each aggregated corpus was divided by the combined corpus, which allows us to see the greatest differences

in word frequency. We reckon that a word with a value of approximately three or higher is specific to a given corpus. Some of these highly frequent words are related to themes, while others are most likely stylistic and related to the genre of the text (literary versus non-fiction, for example).

The methods behind our preliminary findings have raised a number of issues to be worked out and further questions to be asked. For example, both the target corpus and the reference corpus still need to be more rigorously defined. The target corpus was generated from a list of the URNs for texts on the list gleaned from a manual review of the literary histories. This list is problematic because the literary histories often make only tentative associations between—or outright problematize the relationship between—a given text or author and the term "national romanticism." We hypothesize that textbooks used in Norwegian instruction at the secondary school level would be a more appropriate source because they tend to more explicitly link a given text to a given period or style, and we plan to construct a new corpus derived from a manual review of these textbooks. Another reason for using textbooks is that they are much more widely read than literary history books, and thus more central to the formation of the general understanding of the period.

An equally pressing problem with the reference corpus is that it is randomly generated from all books published in the period 1830–1890, rather than being randomly generated from a list of specifically literary books published in the period 1830–1890. We surmise that the inclusion of genres such as, for example, scientific studies, dictionaries, or instruction manuals undermine the validity of our findings, and we thus plan to revise the reference corpus so that it contains only literary texts defined according to standard literary genres.

References:

Jæger, Henrik. 1896. *Illustreret norsk litteraturhistorie.* Bind II. Kristiania: Hjalmar Biglers forlag.
Tangherlini, Timothy R. and Peter Leonard. 2013. "Trawling in the Sea of the Great Unread: Sub-Corpus Topic Modeling and Humanities Research." *Poetics* 41.6: 725–749.
Wilkens, Matthew. 2016. "Genre, Computation, and the Varieties of Twentieth-Century U.S. Fiction." *CA: Journal of Cultural Analytics* (online open-access)
Witoszek, Nina. 2011. *The Origins of the "Regime of Goodness": Remapping the Cultural History of Norway.* Oslo: Universitetsforlaget.
Aarseth, Asbjørn. 1985. *Romantikken som konstruksjon: tradisjonskritiske studier i nordisk litteraturhistorie.* Bergen: Universitetsforlaget.