# Big data approach to 19th-century Finnish newspaper literature

**Mikko Koho, Agata Dominowska, Elsi Hyttinen, Péter Ivanics, Elizabeth Oakes, Ilona Pikkanen, Leena Tulkki, Risto J. Turunen**

What would Finnish literary history look like if, instead of focusing solely on canonical books, we accounted for the works of fiction published in journals and newspapers as well? Traditionally, research on literary history has focused mainly on the close reading of “classics”, constituting about one per cent of the whole literary field (Moretti 2000), which means that our view of historical literature is unbalanced and that can be regarded as a kind of research bias. We began to tackle the question during the intensive week of Helsinki Digital Humanities Hackathon in May 2017, with the scope limited to one genre of fiction: poems.

The National Library of Finland has made historical archives of Finnish newspapers, magazines and other such printed material openly available on the Internet. All material from the 19th century has already been digitized. We show the preliminary results of computationally extracting poems from the 19th century newspapers and perform an initial analysis of the found poems to demonstrate the approach.

For detecting poetry, we used a supervised machine learning approach. We parsed the OCR-recognized texts from the newspaper archive, quantified the features of the texts, and built a classifier to predict the genre of the texts. The trained support vector machine (SVM) classifier set then to work on previously unseen data and attempted to label the text with the best possible match based on hand-picked training data. Source codes for data processing and the poem classifier model are available online[1].

Inspired by the success of other researchers (e.g. Underwood 2014, Hettinger et al. 2015) in the field of text classification using Support vector Machines, we decided to also use an SVM classifier. Word frequencies are used as the main features, and in addition 4 ad-hoc features are calculated from the blocks of text.

## Results and preliminary analysis

With the classifier developed during the hackathon week we found poems from a total of 18,591 newspapers out of 56,985 Finnish-language newspapers published 1800–1890. An overall precision of 87% was obtained while the recall varied between 23% and 67%. Improving the ratio between these two is one of the methodological problems any further research working with the same material should tackle.

[1] https://github.com/dhh17/categories_norms_genres/

In their genre-detecting project Ted Underwood et al. (2014) reached precision rates of poetry that were similar to us, varying between 89.7 % and 90.6%; in the course of their project the recall rates reached over 90 %. However, even with these excellent rations Underwood points out that maximizing recall poses a bigger challenge to all genre-detection projects (2014: 7).

We tried out analysis methods with a corpus of approximately 100 poems. For the analysis, structural topic modelling was used, which yielded promising results – the topics seemed like plausible poem themes: war, love, and Midsummer, among others.

The keyness method, pertaining to corpus linguistics, based on relative word frequencies in the sub-corpus of newspaper poems to relative word frequencies in the whole newspaper corpus, has also proven fruitful. Keywords confirm some of the hypotheses we had based on reading individual poems and previous research: for example, the words connected to nationalism, such as “Finland”, “fatherland” and “birth land”, are indeed overrepresented in the poems compared to the whole newspaper corpus. We also observed that religious words (such as “God”, “sin” and “mercy”) are used much more often in poems than they are in other genres during the whole nineteenth century. Especially the keywords of the 1820s are dominated by religious vocabulary. However, during the following decades, religious poetry loses its hegemony, for the words that distinguish poems from other newspaper texts become more diverse.

Even these preliminary attempts at identifying and analyzing the genres within the archive, with almost 20 000 identified poetry text blocks from the period 1800–1890 have demonstrated that studying works of fiction found in newspapers is a task worth undertaking. Our results suggest that a data-rich history of Finnish newspaper literature is an attainable goal in time, and we hope to be able to pursue it further.

## Bibliography

HETTINGER, L., BECKER, M., REGER, I., JANNIDIS, F. and HOTHO, A., 2015. Genre classification on German novels, *Database and Expert Systems Applications (DEXA), 2015 26th International Workshop on* 2015, IEEE, pp. 249-253.

MORETTI, F., 2000. Conjectures on World Literature, *New Left Review* Jan/Feb 2000 (1), pp. 54-68.

UNDERWOOD, Ted, 2014. *Understanding Genre in a Collection of a Million Volumes*. Interim Performance Report Digital Humanities Start-Up Grant, Award HD5178713.