Dominowska, Agata; Hyttinen, Elsi; Ivanics, Péter; Koho, Mikko; Pikkanen, Ilona; Turunen, Risto

Hiding in Plain Sight: Poetry in Newspapers and How to Approach It

Published in:
HUMAN IT: TIDSKRIFT FÖR STUDIER AV IT UR ETT HUMANVETENSKAPLIGT PERSPEKTIV

Published: 02/07/2019

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Dominowska, A., Hyttinen, E., Ivanics, P., Koho, M., Pikkanen, I., & Turunen, R. (2019). Hiding in Plain Sight: Poetry in Newspapers and How to Approach It. *HUMAN IT: TIDSKRIFT FÖR STUDIER AV IT UR ETT HUMANVETENSKAPLIGT PERSPEKTIV*, 145-171. https://humanit.hb.se/article/view/594

Hiding in Plain Sight: Poetry in Newspapers and How to Approach it

Agata Dominowska,[a] Elsi Hyttinen,[b] Peter Ivanics,[a] Mikko Koho,[c] Ilona Pikkanen,[d] Risto J. Turunen[e] (in alphabetical order)

[a] University of Helsinki, [b] University of Turku, [c] Aalto University, [d] Finnish Literature Society, [e]University of Tampere

In this paper, we describe a computational method to detect poems in the digitised historical newspaper archive of the National Library of Finland, elaborate on the method's strengths and weaknesses, and discuss how digital approaches can be developed further in future research to enhance our understanding of the wide variety of poems published in newspapers. The lack of metadata that would denote content structure and content type posed the biggest challenge for the supervised machine learning approach chosen for the project, but in the final dataset the poetry content had risen to a total of 18,591 text blocks with poetic content, with overall precision rates verging on 90 percent. We argue that even these preliminary results demonstrate that studying poetry published in newspapers is a task worth undertaking. Moreover, the corpus extracted can already enable content-oriented research and we discuss some methods enabling this in the article. Finally, our paper suggests that a data-rich history of Finnish newspaper literature is an attainable goal in time, and it has potential for challenging the current understanding of the Finnish literary past. The paper is oriented towards literary scholarship, by applying computational distant reading to a large literary corpus.

Keywords: literary history, national literature, poetry, Finland, genre detection, nineteenth century

Anyone who has ever skimmed through a newspaper published in the nineteenth century is left with an impression of a wide variety of textual content: besides domestic and international news articles, there are political pamphlets, weather forecasts, obituaries, shipping news, prices of firewood, grain and bicycles, announcements of published books and lists of travellers. There are feuilletons, short stories, plays and poems: besides transmitting information, the newspaper was also an easily available and widely circulating platform for publishing fictional texts of different kinds. However, no one knows how much or what kind of fiction was published in newspapers during the nineteenth century. This observation was the point of departure for a short research project that set out to detect fictional content in digitised historical newspaper archive of the *National Library of Finland*, conducted during the intensive week of *Helsinki Digital Humanities Hackathon* in May 2017. This article discusses the potential directions and implications of the project.

Our overarching question was, what would literary history look like if we did not focus solely on canonical works published in a book format, but considered fictional texts published in journals and newspapers as well? What kind of literature is 'hiding', as it were, on the pages of newspapers? This inquiry is informed by the new paradigm of literary history, related to the question of scale, emergent since the 2000s. Franco Moretti famously argued that literary history relies on detailed textual analysis or 'close reading' of one per cent – the canonical one – of the whole literary field. These rare and exceptional works are, however, by definition not representative of any textual culture as a whole (Moretti 2000, 55; Moretti 2007). The more recent approaches in the field are firmly grounded in computational methods and literary modelling; at the same time, they are cognizant of the limitations of big data, particularly with regard to historical contexts of literature (Jockers 2013; Bode 2017; Kumar & Tucker 2017; Piper 2017a; 2017b). We were inspired by such data-rich literary studies and motivated our short project with the aforementioned big literary historical question, but for pragmatic reasons we

decided to focus on the automated identification of poems in the newspaper corpus and to use a supervised machine learning approach.

The *National Library of Finland* has an ongoing project aimed at making historical archives of Finnish newspapers, magazines and other printed materials freely available on the Internet. The entirety of nineteenth-century texts has already been digitised. The newest set of material from the period of 1910–1920 was released just as we were planning the hackathon. In total, the corpus is comprised of nearly 3,000 000 pages. The historical newspaper archive offers a convenient web interface that supports word searches within the digitised corpus – the user can obtain the list of newspapers, which contain the search terms in a matter of seconds. The archive thus provides an invaluable research tool for many disciplines among the humanities and enables much more extensive research into history and historical social sciences than what was possible using the older, microfilmed newspaper archive.

It is also easy to find particular works of fiction in the archive by searching for an author, a pen name or a title, but the present user interface does not enable searching for and extracting all occurrences of a genre. However, this factor alone does not explain the lack of literary scholarly interest in fictional works published in newspapers and journals, with few notable exceptions, like Katharine Bode's project on literature published in Australian newspapers (2012). At least since feminists started to question male-centred literary histories in the 1960s and 70s, it has been one of the guiding principles of literary history writing to try to take into account forgotten texts and authors; to go beyond the literary landscape of "Austen peaks, the Bronte cliffs, the Eliot range and the Woolf hills" (Showalter 1977, *vii*). However, even when going beyond the canonical works or questioning canon formation, literary scholars usually understand literature as something published as books.

As a contrast, we argue that this vast material of texts should be of immediate scholarly interest to anyone whose work has to do with nineteenth-century Finnish literary culture. Indeed, the book historian Kai

Häggman has pointed out that, "[T]he grand narrative of Finnish publishing history has proceeded from one great Finnish novel to the next, even though the most popular published items in the nineteenth century were in fact religious booklets, thin chapbooks of songs, and ragged newspapers" (2008, 17–18). Exploring and studying literary texts published in newspapers and periodicals would thus opened up a completely new spectrum of widely distributed and in many cases locally produced literary culture (Bode 2012). Moreover, considering the newspaper format would also highlight scholarly questions related to quantity. The latest comprehensive description of Finnish literary history, *Suomen kirjallisuushistoria 1–3* [The Finnish Literary History 1–3] (Varpio & Koskela 1999), consciously steers away from canons and *great men, great works*-approach and asks how literature is expressive of the ideologies and literary ideals of its time. However, it does not offer readers numerical data to back the characterization of literary styles from various periods. The present paper steers towards literary scholarship that takes questions related to both quantity and quality – distant and close reading – into account and merges them into a new way of writing literary histories.

Analysis

Historical newspaper archive

The digital newspaper archive of the *National Library of Finland* is an open access repository of whole text materials published in Finnish and in Swedish, not collected to satisfy any particular need (as opposed to, for instance, various digitised collections of literary classics) and not selected based on any criteria other than the newspaper medium and the date of publication. The database contains all newspapers published in Finland from the earliest years of Finnish journalism to the first decade of the country's independence, in 1920s (see figures 1 and 2). The scope of this

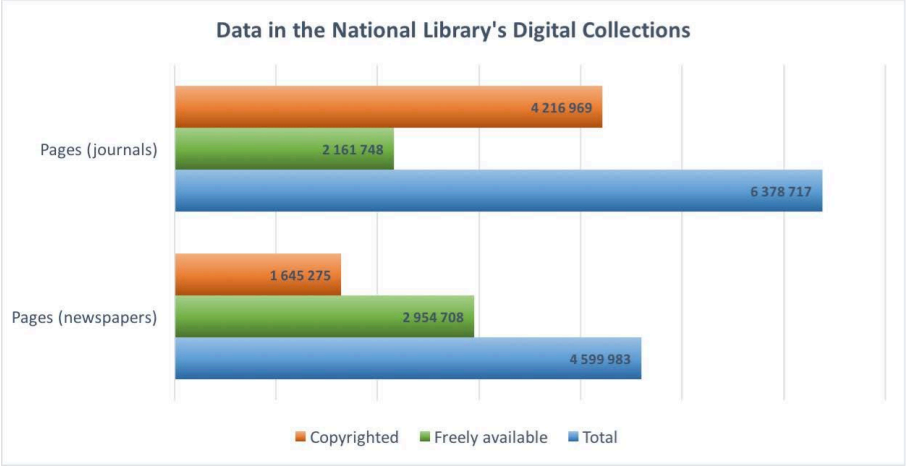

Figure 1: Openly available data in the National Library's Digital Collections. Source: DIGI – National Library's Digital Collections

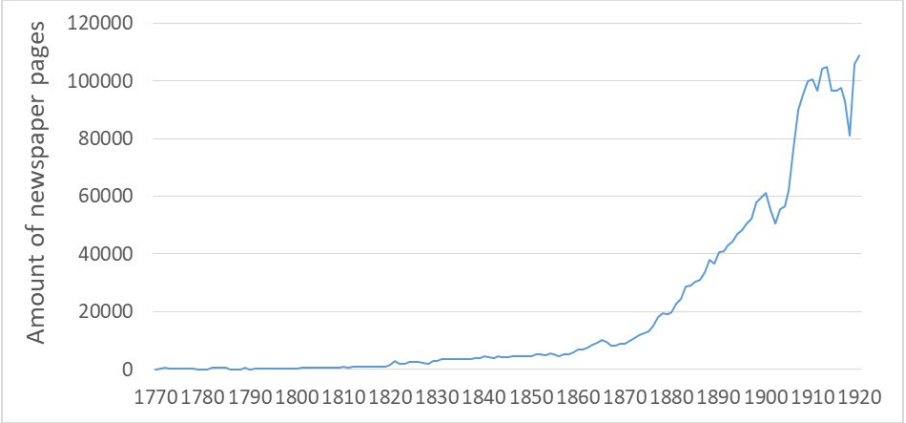

Figure 2: The number of digitised newspaper pages in the Newspaper Archive. Source: DIGI – National Library's *Digital Collections*

study is limited to newspaper content published in the 19th century in the Finnish language. Finnish 19th-century press existed in both

languages, Finnish and Swedish, but each language would require a different algorithm for genre detection.

The archive provides a dataset that is particularly apt for studies using data mining techniques. The data is offered as offline bundles until the year 1910,[1] and it contains information in different formats (Pääkkönen *et al*. 2016), out of which the ALTO XML (Library of Congress 2016) has proven to be the most useful for the purposes of this study. These data files enabled us not only to access the contents of the newspaper pages, but also to obtain information about their layout and positioning on the pages. The content of the newspapers is stored in a nested structure consisting of text blocks, text lines and individual words. These have been automatically generated from newspapers using optical character recognition (OCR). The OCR accuracy varies greatly within the corpus.

This variance may have resulted from differences in, for example, typeface, quality of the original copy, page layout, and many aspects of the digitisation process (Kettunen *et al*. 2016). Naturally, it also affects the usability of the corpus for processing with computational methods. Further challenges stem from data structure, such as inconsistent order and varying size of text blocks and incorrect segmentation of logical groups of textual content, for instance the delimitation of articles.

The newspaper archive has been categorised partly into an article index. This categorisation was performed manually between 1890–1909 as a joint effort of many scholarly and scientific associations of the time, and only covers material published before 1890.[2] The index seems to be mainly based on the topic of the text instead of the text type, *i.e.* its genre, with some exceptions, such as poetry. However, the index is largely incomplete, and most of the poems we identified using the classifier algorithm were not listed. Nevertheless, this historical index provided a valuable starting point to our computational genre detection, as we were able to use material listed in it as training data for the classifier.

Algorithmic Genre Classification

For detecting poetry, we chose 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 in order to predict whether a text belongs to the poetry genre or not. It should be pointed out that at this stage, we were developing the algorithm to recognize texts that meet common criteria characteristic of poetry; the assessment of literary value and artistic quality would have to wait until later stages of the project. The support vector machine (SVM) classifier, trained on hand-picked training data, was then set to work on previously unseen data and attempted to label the text with the best possible match. All source code for data processing and the poem classifier are available online.[3]

The OCR-recognized texts hold only the following structure: words, lines, blocks of text (usually text divided into paragraphs), pages, and issues. What is missing is the content structure, *i.e.* the metadata about the place where one article, advertisement or another type of independent content item begins and ends. Content structure is also a prerequisite for other key metadata, namely the content type, *i.e.* the genre of each item (for instance the classification of a given news piece as announcement, advertisement, short story, play, poem, *etc.*)

There is plenty of existing research in text classification in general (Aggarwal *et al.*, 2012), and the more specific topic of classifying texts based on their genre has a long tradition (Kessler *et al.*, 1997). Among others, Stamatatos *et al.* (2000) have applied occurrence frequencies of common words to text genre detection. However, to the best of our knowledge, there has not been any research on automatic genre detection of Finnish texts. There is existing research, in which algorithms have been used to separate poetry from other genres (Tizhoosh *et al.* 2008; Jamal *et al.* 2012; Underwood 2014). Tizhoosh *et al.* (2008) used supervised learning with multiple features to distinguish poetry content from

prose. Tizhoosh *et al.*'s corpus is relatively small (850 documents), but within the scope of our present project, their method of adding several features, such as word frequencies, rhymes and shape, to a classifier is of particular interest as it iteratively increases the accuracy of the model.

We decided to use SVM classifier to classify text items into poems and other content ("non-poems"), based on the success of the SVM in previous genre classification research (Underwood 2014, Hettinger *et al.* 2015). Although in a previous study (Underwood 2014) logistic regression models outperformed SVMs, it was still considered likely that tuning of SVM parameters would lead to a better performing model. We also tried logistic regression, which did provide us with the highest accuracy at a certain stage during the project. However, when we added features that further improved the classification accuracy, the SVM consistently performed better.

To make the unstructured data usable for supervised machine learning, we needed to create a vector representation of each piece of text. Initial vectorization consisted only of word frequencies, normalized with tf-idf. The words were not lemmatized, but used in their written form. Eventually, the amount of word frequency features was limited to the 25 400 most common words, since this approach resulted in the best training accuracy.

Afterwards, more features were tested and added. For this purpose, we identified the features of poems that a human eye would look for. The features were chosen by manually exploring the corpus for frequently occurring visible differences between poems and other texts. We observed that the poems were, *e.g.*, usually consisting of shorter rows, each row often starting with a capital letter, and noticed that adding these features actually improved the cross-validation accuracy. We ended up using four features in addition to the word frequencies, which improved the classification precision. These were the:

1. Average text row length,
2. Number of one-word rows,
3. Number of dots,
4. Number of rows starting with a capital letter.

In Finnish-language poetry, words are often shortened by using apostrophes; therefore, a high number of apostrophes seemed like a useful feature to add to the classifier. However, it soon became evident that in most cases, the OCR tool is unable to identify them correctly and we had to desist from considering apostrophes as a feature.

The next step was to gather training data. The initial selection of texts was done with the help of the aforementioned article index from the 1890s. In the first iteration, approximately 400 text blocks with poetry content (corresponding to a hundred poems), were collected manually. Once the poetry content was identified, the corresponding newspaper, page number and text block identifiers were listed on a spreadsheet and fed into the machine learning algorithm. All other content on those pages was then used as training data for the non-poem class. Five percent of the training data acquired in this way consisted of poetry, among a total of 120 pages. The ratio of poetry to non-poetry in this particular training dataset was considerably higher than the proportion of poetry in the entire corpus. We considered this finding to be a good sign; with too little poetry content, the model would learn to label everything as non-poetry, since this would produce the least amount of classification errors. Once the classifier learned to identify poetry, we were able to increase the amount of training data to improve the accuracy of the classifier.

The data was then used to adjust the classifier. From early on we were applying an exhaustive search over a range of possible hyper-parameter values of the classifier. The classifier uses a linear kernel, and is implemented with stochastic gradient descent training. The best performing combination of hyper-parameters were selected, *i.e.* the ones with the

best 3-fold cross-validation accuracy. The hyper-parameters also dictate the type of the linear model (SVM, logistic regression, *etc*.) through the loss function. The training accuracy increased gradually as more training data was gathered and more hyper-parameters added to the grid search, resulting in training accuracy of up to 92 percent. Unfortunately, but not surprisingly, we observed that the classifier was unable to properly distinguish poetry from various kinds of lists (for example name lists, schedules, weather forecasts) and of advertisements, classifying them all as poems. We tried to overcome this obstacle by running the data through an improved classifier that had been trained with an increased number of lists in the non-poem data. The model, however, did not perform better on the whole corpus, but instead resulted in lower precision. The underlying reason as to why enriching non-poem training data with additional lists did not help could be that the non-poem class was comprised of various genres grouped into one. It appears that the lists skewed the average feature values of the non-poem class away from what would be typical of prose, and closer to values typical for poems. As a consequence, the model became less accurate in making the distinction between poetry and prose.

The Varying Landscape of Newspaper Poetry

With the classifier developed during the hackathon week, we found poems in a total of 18,591 newspapers out of 56,985 Finnish-language newspapers published 1800–1890. In the final dataset, the overall precision is a good 87 percent (that is, 13 percent of the hits that the algorithm found are not poems), based on a test sample of 324 identified text blocks, out of which 283 had actual poetry content (see table 1). The false positives consist of different kind of lists and, to a smaller extent, of short prose segments. At the same time, however, the recall rate (the number of published poems caught by the algorithm) remained relatively low (see table 2) with significant variation from decade to decade.

Improving the ratio between these two is one of the methodological problems any further research working with the same material should tackle.

PRECISION Chart 3	1830	1840	1850	1860	1870	1875	1880	1885	1890	
Amount of pages	2950	4408	4445	6808	8767	13270	17952	30415	40804	129819
Amount of identified and checked text blocks	8	11	11	17	21	33	45	76	102	324
Poetry content	7	11	10	15	21	28	43	58	90	283
Precision rate %	**87,5**	**100**	**91**	**88**	**100**	**85**	**95,5**	**76,3**	**88**	**87**

Table 1: Precision on a sample of years

RECALL Chart 4	Dec 1829	Dec 1849	Dec 1859	Dec 1869	1.-5.12.1884	IN TOTO
The amount of published newspapers	2	2	7	9	24	44
Poems found manually	6	0	12	48	5	71
Poems found by the classifier	4	0	4	11	3	22
Recall rate %	**67**	**100**	**33**	**23**	**60**	**31**

Table 2: Recall on sample of months

In their genre detecting research, Underwood *et al.* (2014) reached precision rates of poetry that were similar to our results, varying between 89,7 percent and 90,6 percent; in the course of their (much longer) project the recall rates reached over 90 percent. However, even with these

excellent ratios, the authors point out that maximizing recall poses a bigger challenge to all genre detection projects (2014, 7).

Moreover, Underwood *et al*. interestingly reflect on the scholarly implications of the trade-off between precision and recall in the context of literary scholarly research. A strict definition of a genre maximizes precision, whereas a more lenient definition increases recall and hence also the amount of false hits. By tuning the classifiers, it would be possible to find a balance between the two, but what is the right balance? Literary scholars do not care about precision and recall in equal measure; they are used to working with 'pure' corpora, *i.e*. very small collections that implicitly exclude the majority of published materials. In other words, these corpora have low recall, but very high precision (Underwood 2014, 31).

An additional challenge, when it comes to balancing between varying degrees of precision and recall in the context of newspaper literature, is the potential diversity of literary production in a medium that, even with possible censorship and some degree of editorial gatekeeping, might have allowed explicit and implicit literary experimentation. In other words, we are not dealing with the rare and exceptional works of the canon, which literary scholars know by heart, but with the 'banal, everyday' (Moretti 2007, 3) forms of literature. Do these forms of literary production, and more particularly of poetry, look like their canonical counterparts published on the pages of monographs? To what degree do our presuppositions affect the training data and, consequently, the kind of texts we find?

This wider question became topical when the reasons for the particularly low recall rate of 1869 (23 %) were surveyed. The noise could partly be explained by translated poetry, which the algorithm did not identify with the same efficiency as Finnish source texts. When translations were removed, the recall rate of 1869 rose to 52 percent. The problem might have been caused by our initial training data composition; translated poetry was absent from both the index and the hand-picked training

data. The amount of translated poems has been evaluated as "exceptionally low" in previous research which quantified manually newspaper poetry published in one town during the late nineteenth century (Lassila 1979, 64). As a contrast, our project signals that there might be more translated poetry published in newspapers than previously presumed – probably with great quantitative variation depending on the region and the profile of each newspaper. Even so, interesting questions remain also as to the form and content of translated poetry. Are their vocabularies and other attributes so different from the Finnish poems in the original that the classifier failed to detect them?

Moreover, during the tests for precision and recall it was also noticed that the classifier needs further improvement to overcome the problems posed by the lack of metadata about content type (where one generic entity ends and the next begins), as the classifier works on the level of texts blocks only. When the results were investigated in more detail, it turned out that a hit – "a poem"– might in fact be either a complete poem or any of the following options:

a) A disconnected, individual text block with poetry content,
b) Several but not all text blocks of the same poem,
c) Several text blocks of different poems (on the same newspaper page).

In other words, we need to think of ways to improve also the completeness of the detected poetry content; the algorithm should identify all text blocks of the same poem, and not just some of them. The following ideas were considered as possible next steps to improve the recall rates and the completeness of poems recognised:

a) Could we utilize the morphological analysis of identified poems to improve recall?
b) Could the algorithm be developed to retrieve and inspect the surrounding text blocks of the exact same width as a text block that had already been identified as a poem, to improve the completeness?
c) Could we use the segmentation information (METS files) for the same purpose?
d) Could we use the information on element positioning on the page, contained in the metadata, to group text blocks into articles?

The goal of the present project was to find as many of the poems published in newspapers as possible (high recall), and at the same time say something sensible about the amount of newspaper poetry (high precision) and, in the long run, also about the content of it (high precision). During the project, we have, with the help of the algorithm – even with their low recall rate – identified a number of formerly unknown poems. Compared to the starting point, the hand-picked index of 4000 poems, our dataset is more representative of the poetry published in newspapers. In the longer run, the study into literary texts published in newspapers and journals, combined for example with bibliometric analyses, has potential to shed light on the literary landscape from the quantitative perspective, promising a more holistic picture of the textual culture from which the canonical works stem, and vice versa. Moreover, although Underwood doubts the literary scholars' willingness to work with a collection of poetry that contains up to 10 percent noise (2014, 31), these preliminary precision rates would allow us to make inferences about qualitative questions too, with the margin of error that mostly revolves around 13–15 percent.

Discussion

From the point of view of literary scholarship, identifying a corpus is not an interesting result *per se*; establishing the research material is where the real questions only begin. During the hackathon week, we only had time to test the potential of some well-known computational methods – topic modelling, keyness and morphological analysis – for content analysis. Close readings, contextualising and comparative literary analysis vis-à-vis the existing scholarship on the Finnish literary history remains to be done at a later stage.

We started with topic modelling, with a set of training data consisting of 100 poems (Koho *et al*. 2017). The algorithm returned results both surprising and not so surprising: topics containing bellicose expressions around the Crimean War (1853–1856) make sense immediately, and so do the nationalistic tones, but most of the word groups had no such self-evident explanation and were more difficult to label. Why were there a number of poems on the pleasures of having a cup of coffee? A more promising approach was offered by the keyness analysis. It pertains to corpus linguistics and is based on relative word frequencies. To count as a keyword, the word needs to be significantly more frequent in the studied corpus than in the reference corpus (Baker 2006, 125–126). In our case, we compared relative word frequencies in the sub-corpus of newspaper poems to relative word frequencies in the whole newspaper corpus.[4]

These search terms were used to quantify Finnish nationalism in the newspapers: "Suomi", "suomikin", "suome?", "suome??", "suomessa", "suomessa*", "suomemme", "synnyinmaa*" and "isänmaa*". The search produced 4 260 hits in 848 684 words in the newspaper poetry. The search produced 688 233 hits in 338 331 078 words in the whole newspaper corpus.

Not entirely surprising in the nineteenth-century European context in which literature was used to imagine new nations (see *e.g.* Anderson

1983), words connected to nationalism, such as "Finland", "fatherland" and "birth land", are overrepresented in the poems compared to the whole newspaper corpus (see figure 3). Moreover, the amount of nationalist content in poems seems to increase by the decade towards the turn of the century, whereas in the whole corpus nationalism has a more static, but a less notable role. We also observed that religious words (such as "God", "sin" and "mercy") are used much more often in poems than they are in the rest of the corpus during the whole nineteenth century. Especially the 1820s are dominated by religious vocabulary: during the following decades, religious poetry loses its hegemony, as the words that distinguish poems from other newspaper texts become more diverse.[5]

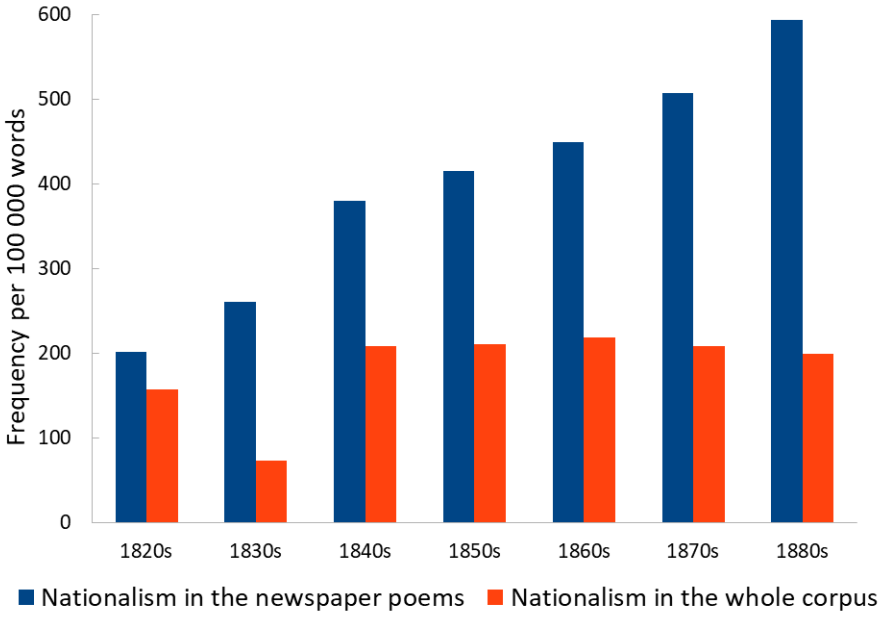

Figure 3. The relative frequency of nationalist words by decade.

On the whole, the observation of nationalism and religion as dominant themes of poetry is in no way surprising in the framework of

nineteenth century (textual) culture. Nevertheless, the question of the quantitative variation of these themes in the newspaper poetry versus the whole newspaper corpus is potentially interesting, as is the temporal change indicated by the preliminary results. Far too often, word frequency charts presented in digital humanities are genre-blind, neglecting the distribution of different genres inside a given corpus. Here, in order to understand these diversifications, our next step would be to enrich the newspaper metadata with political affiliation, and to analyse how the master ideologies of the long nineteenth century – especially nationalism, liberalism and socialism – affect the language of poetry.

While the keyness method combined with close reading can deliver promising results, it obviously has weaknesses too. In practical terms, OCR-errors and widely circulated poems, copied from one newspaper to the next, may skew the results. From a theoretical perspective, keyness operates on a single word level, thus ignoring relations between words. We have evidence that nationalism is an expanding theme within the poetry genre, but how does nationalism in poetry change in the course of the nineteenth century? Keyness does not give any information on the context-bound meanings of words; rather, it shows which words should be analysed more carefully with supportive methods, such as morphological analysis.

For detecting variance in terms of temporality, location, and individual style, morphological analysis can provide interesting possibilities. For a rough grain analysis of style, identifying trends in use of specific linguistic features in a genre across time, computers may perform better than humans. Being a condensed genre, poetry encodes meanings simultaneously at multiple textual levels with, generally speaking, greater rigor than longer form genres, such as novels. While poetry imposes some formal restraints, it also affords greater freedom of choice in word class and morphological structures. Words need not occur in the usual order, and the rules of grammar are generally relaxed. Word class proportions

need not mirror either speech or other genres of writing. For example, a poem might plausibly contain a higher proportion of adverbs to verbs than would be tolerated in other genres as a matter of style. A large-scale analysis of the morphological structure of POS-tagged inflected words included in the poetry data – their stems, root words, prefixes, suffixes, and the parts of speech they represent – may uncover possible trends in the way the poems were written and the general features of the genre, such as, hypothetically, an increase in the use of verbs (poems becoming more active) in a particular decade or a distinct diachronic pattern of use of interjections. Complementing keyness and other methods that approach the corpus as a whole, this method would allow a further insight into the inner variation within the corpus and bring focus to regional differences between texts, possibly highlighting the spatial and temporal spread of regional or dialectal features within the genre of poetry.

One of the tangible and interesting preliminary results offered by our project is that besides artistically ambitious poems, newspapers feature context-dependent occasional poetry (Wilson 2012): verses that celebrate engagements, graduations and other festivities, poems that commemorate lost persons and that address always topical themes such as alcohol and coffee consumption and dog taxation. In other words, occasional poetry published in newspapers is one of the 'hiding' literary forms that a data mining approach, such as the described in the present paper, is able to unravel. Using traditional humanist methods, this formerly discarded form of poetry has recently been studied for example in the framework of 'self-taught writers' (Laitinen & Mikkola 2013). These studies concentrate on occasional poetry published in broadsheets or other forms of small prints. However, it is probable that newspapers will turn out to be the nineteenth-century treasure trove for such poems, although the sheer volume of newspaper poetry makes it impossible to grasp the phenomenon in its entirety without computational solutions, as was already noted in the Finnish press history of the 1970s (Lassila 1979, 73; see also

Laitinen 1981). Overall, beyond the literary historical research interests of our project group, such computationally identified corpus of newspaper poetry would provide, for example, a fresh and ample research material for scholars studying uneducated writers, or, material for scholars interested in the intertwined nature of oral and literary cultures in the long nineteenth century.

Conclusion

In this article, using the example of poetry detection, we have presented some of the computational methods that would enable a new, broader and more inclusive history of the Finnish literature, a history that would take into account the "banal" and the "everyday" of literary texts, and also address the much wider question of textual availability of literature and its circulatory (or in the newspaper context even viral) quality (Bode 2017; Nivala *et al*. 2018). As the article describes, we have not yet been able to create an algorithm that would work well enough to enable near 100 percent accurate detection of poetry in the *National Library of Finland*'s digital newspaper archive, and more work needs to be done particularly in the matter of recall. However, even these preliminary stages with almost 20 000 identified poetry text blocks from the period 1800–1890 have demonstrated that studying works of fiction published in newspapers, and particularly poems, is a task worth undertaking. Moreover, the corpus extracted to this end, with overall precision rates verging on 90 percent, can already enable content-oriented research. Although the results and the conclusions must remain open-ended at this stage, being more indications for further research rather than concrete outcomes, the present paper does suggest that a data-rich history of Finnish newspaper literature is an attainable goal in time, and it has potential for challenging the current understanding of the Finnish literary past.

Both practical and theoretical arguments have been raised against the use of such quantitative methods in literary studies (*e.g*. Hoover 2008).

Particularly interesting in the framework of literary scholarship is Katherine Bode's recent explanation for the sharp polarity between close and distant readers; according to her, the reason is the abstract and ahistorical way many distant readers approach their data. As Bode argues, the neglect of historical, contextual insight is not a result of applying computational methods to literary history, but rather it is inherited from the New Criticism (2017, 2). At the same time, the literary scholar Andrew Piper reminds us that the antipode to (computational) measurement is not subtlety or complexity, but personal authority: "Measurement replaces charisma as the guiding vehicle of generalization" (2017b, 654). The present paper demonstrates that there is no reason to believe that a distant reading approach, aimed at charting the entirety of a given literary field during a particular time period, should render close reading invalid, or that data acquired in this way should be incompatible with the inherent complexity of literature. However, newspapers were a democratizing media in the nineteenth century, and perhaps one can make the same claim about digital humanities: their methods make literary historical knowledge not only cumulative but also conglomerative (*ibid.*, 655).

It has also been pointed out that, in the era of diminishing resources, there is a real danger that digitisation projects will favour canonical materials, which shall further marginalize less esteemed archives and research topics (Guldi & Armitage 2012, 113). In this respect, the role of the digital newspaper is twofold. Firstly, it is a well-established research resource. The importance of the newspaper medium in the development of the 19th century bourgeois public sphere and, consequently, of the nation-state, is widely recognised, thus historians and social scientists routinely use digitised newspapers in their research. On the other hand, in regard to literary history, the newspaper collection constitutes an archive of the past that is forgotten and non-canonised. In this paper, we employed both computational methods and those typical of the humanities in order to reflect and account for this dual nature. The outcome was the

first glimpse into the marginal history of everyday literature preserved inside the reservoir of this well-known textual corpus, hiding, as it were, in plain sight.

Agata Dominowska is currently pursuing a Master's degree in Finnish Language and Culture, with a background in applied linguistics and English philology. Since 2014, she has been working as a research assistant at the University of Helsinki in a project combining digital methods and historical sociolinguistics.
Contact: agata.dominowska@helsinki.fi

Doctor Elsi Hyttinen has a PhD in Finnish literature (2012, University of Turku, Finland). Her special fields of interest are queer literary history and posthumanist theory. Hyttinen is currently writing a monograph on depictions of Finnish-American migrancy in early 20th century Finnish prose.
Contact: elshyt@utu.fi

Péter Ivanics studies at the University of Helsinki on the master degree of Computer Science. Prensently he is also working at a local startup as an iOS Software Engineer. His research interests include Data Mining, which is also the topic of Ivanics Master's thesis.
Contact: peter.ivanics@helsinki.fi

Mikko Koho is a doctoral candidate at the Semantic Computing Research Group in the Department of Computer Science at the Aalto University School of Science (Helsinki, Finland). Koho´s main research interests are Linked Data and Knowledge Discovery, and applying these to the interdisciplinary field of Digital Humanities.
Contact: mikko.koho@aalto.fi

Doctor Ilona Pikkanen has a PhD in history (2013, University of Tampere, Finland). Her current research focuses on long-term comparative historiography, historical fiction and literary history with focus on questions that can best be solved in the framework of digital humanities. Pikkanen works at the Research Department of the Finnish Literature Society, Helsinki.
Contact: ilona.pikkanen@finlit.fi

Risto J. Turunen is a doctoral candidate at the University of Tampere. He is currently writing his PhD thesis on the language of Finnish socialism, 1895–1918, combining approaches from labour history, conceptual history and digital humanities.
Contact: risto.turunen@tuni.fi

Notes

1. The *National Library of Finland* (2017), The open data exports of digitised newspapers and journals, 2017. The *National Library of Finland*. <https://digi.kansalliskirjasto.fi/opendata/submit> [2019-04-11]
2. The newspaper archive partly categorised into an article index <http://blogs.helsinki.fi/scriptaselecta/2017/07/13/digitaalisten-aineistojen-artikkelihakemiston-historiasta/> [2019-05-20]
3. Source code for data processing and the poem classifier available online <https://github.com/dhh17/categories_norms_genres/> [2019-04-11]
4. The keyword analysis requires statistical metrics. For the discussion on the best method for calculating keyness, see Brezina & Meyerhoff (2014); Gabrielatos & Marchi (2012); Pojanapunya & Wattson Todd (2016); we used the log-likelihood test
5. The exact statistics for keywords of each decade: <http://github.com/dhh17/categories_norms_genres/tree/master/keyword> [2019-05-20]

References

AGGARWAL, CHARU C. & CHENGXIANG ZHAI (2012). "A Survey of Text Classification Algorithms." *Mining Text Data*. Eds. Charu C. Aggarwal & ChengXiang Zhai. Boston, MA: Springer.

ANDERSON, BENEDICT (1983). *Imagined Communities. Reflections on the Origin and Spread of Nationalism*. London: Verso.

BAKER, PAUL (2006). *Using Corpora in Discourse Analysis*. London: Continuum.

BODE, KATHERINE (2012). *Reading by Numbers: Recalibrating the Literary Field*. London: Anthem Press.

BODE, KATHERINE (2017). "The Equivalence of "Close" and "Distant" Reading; Or, Toward a New Object for Data-Rich Literary History." *Modern Language Quarterly* 78.1: 77–106.

BREZINA, VACLAV & MIRIAM MEYERHOFF (2014). "Significant or Random? – A Critical Review of Sociolinguistic Generalisations Based on Large Corpora. " *International Journal of Corpus Linguistics* 19.1: 1–28. DOI: 10.1075/ijcl.19.1.01bre

GABRIELATOS, COSTAS & ANNA MARCHI (2012). Keyness: Appropriate Metrics and Practical Issues. CADS International Conference 2012. *Corpus-assisted Discourse Studies: More than the Sum of Discourse Analysis and Computing?,* 13–14 September, University of Bologna, Italy. Also available as: <http://repository.edgehill.ac.uk/4196/> [2018-12-08]

GULDI, JO & DAVID ARMITAGE (2014).*The History Manifesto.* Cambridge: Cambridge University Press. Also available as:

<https://www.cambridge.org/core/services/aop-file-manager/file/57594fd0fab864a459dc7785/historymanifesto-2Oct2014.pdf> [2019-05-20]

HETTINGER, LENA ET AL. (2015). "Genre Classification on German Novels." *Database and Expert Systems Applications* (DEXA), 2015 26th International Workshop on 2015, IEEE: 249–253.

HOOVER, DAVID L. (2008). "Quantitative Analysis and Literary Studies." *A Companion to Digital Literary Studies*. Eds. Susan Schreibman & Ray Siemens. Oxford: Blackwell. Also available as: <www.digitalhumanities.org/companionDLS/> [2019-05-20]

HÄGGMAN, KAI (2008). *Paras Tawara maailmassa. Suomalainen kustannustoiminta 1800-luvulta 2000-luvulle.* Helsinki: Otava.

JAMAL, NORAINI, MOHD MASNIZAH & NOAH SHAHRUL AZMAN (2012). "Poetry Classification Using Support Vector Machines." *Journal of Computer Science* 8.9: 1441–1446.

JOCKERS, MATTHEW L. (2013). *Macroanalysis: Digital Methods and Literary History*. Urbana, Chicago, Springfield: University of Illinois Press.

KESSLER, BRETT, GEOFFREY NUNBERG & HINRICH SCHÜTZE (1997). "Automatic Detection of Text Genre." *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics. 32–38.

KETTUNEN, KIMMO, TUULA PÄÄKKÖNEN & MIKA KOISTINEN (2016). *Between Diachrony and Synchrony: Evaluation of Lexical Quality of a Digitized Historical Finnish Newspaper and Journal Collection with Morphological Analyzers*. Baltic HLT.

KOHO, MIKKO ET AL. (2017). Rewriting Literary History with Big Data. Poster at the *Helsinki Digital Humanities Hackathon* 2017. Also available as: <http://www.finlit.fi/sites/default/files/mediafiles/tutkimus/posterdraft2_1.pdf> [2019-05-20]

KUMAR, KRISHNA & HERBERT F. TUCKER (2017). "Introduction." *New Literary History* 48.4: 609–616.

LAITINEN, KAI (1981). *Suomen kirjallisuuden historia*. Helsinki: Otava.

LAITINEN, LEA & KATI MIKKOLA, EDS. (2013). *Kynällä kyntäjät. Kansan kirjallistuminen 1800-luvun Suomessa. SKST 1370*. Helsinki: SKS.

LASSILA, HELENA (1979). *Runot Oulun sanomalehdistössä 1879–1905. Runoutta ja nurkkaromaaneja – Sanomalehdistö kaunokirjallisuuden julkaisijana ennen vuotta 1917. Eds. Suomen sanomalehdistön historia -projekti*. Helsinki: Suomen Sanomalehdistön Historia-projektin julkaisu 14: 60–74.

LIBRARY OF CONGRESS (2016). *ALTO: Technical Metadata for Layout and Text Objects*. <https://www.loc.gov/standards/alto/> [2019-05-20]

MORETTI, FRANCO (2000). "Conjectures on World Literature." *New Left Review* 2000.1: 54– 68.

MORETTI, FRANCO (2007). *Graphs, Maps, Trees: Abstract Models for Literary History*. London: Verso.

NIVALA, ASKO, HANNU SALMI & JUKKA SARJALA (2018). "History and Virtual Topology: The Nineteenth-Century Press as Material Flow." *Historein* 17.2. DOI: http://dx.doi.org/10.12681/historein.14612

PIPER, ANDREW (2017 a). Data, Data, Data. Why Katherine Bode's New Piece is So Important and Why It Gets So Much Wrong about the Field. Available at: <https://txtlab.org/2017/06/data-data-data-why-catherine-bodes-new-piece-is-so-important-and- why-it-gets-so-much-wrong-about-the-field/> [2019-04-11]

PIPER, ANDREW (2017b). "Think Small: On Literary Modeling." *PMLA* 132.3: 651–658.

POJANAPUNYA, PUNJAPORN & RICHARD WATSON TODD (2016). "Log-likelihood and Odds Ratio: Keyness Statistics for Different Purposes of Keyword Analysis." *Journal of Corpus Linguistics and Linguistic Theory*. Published online: April 2016. DOI: 10.1515/cllt-2015-0030

PÄÄKKÖNEN, TUULA ET AL. (2016). "Exporting Finnish Digitized Historical Newspaper Contents for Offline Use." *D-Lib Magazine* 22.7–8. Also available as: <https://doi.org/10.1045/july2016-paakkonen> [2019-05-20]

SHOWALTER, ELAINE (1977). *A Literature of Their Own: British Women Novelists from Brontë to Lessing*. Princeton, N.J.: Princeton UP.

STATAMATOS, EFSTATHIOS & GEORGE K. KOKKINAKIS (2000). "Text Genre Detection Using Common Word Frequencies." *Proceedings of the 18th Conference on Computational Linguistics* 2: 808–814. Association for Computational Linguistics.

TIZHOOSH, HAMID R., ROZITA DARA & FARHANG SAHBA (2008). "Poetic Features for Poem Recognition: A Comparative Study." *Journal of Pattern Recognition Research* 3.1: 24–39.

UNDERWOOD, TED (2014). Understanding Genre in a Collection of a Million Volumes. Interim Performance Report Digital Humanities Start-Up Grant, Award HD5178713. Also available as: <https://figshare.com/articles/Understanding_Genre_in_a_Collection_of_a_Million_Volumes_Interim_Report/1281251> [2019-05-20]

VARPIO, YRJÖ & LASSE KOSKELA, EDS. (1999). *Suomen kirjallisuushistoria* 1–3. Helsinki: Suomalaisen Kirjallisuuden Seura.

WILSON, STEPHEN (2012). "Poetry and its Occasions. Undoing the Folded Lie." *A Companion to Poetic Genre*. Ed. Erik Martiny. Hoboken, N. J.: Wiley-Blackwell. 490–504.