Digital interfaces of historical newspapers: opportunities, restrictions and recommendations

Eva Pfanzelter, Sarah Oberbichler, Jani Marjanen, Pierre-Carl Langlais, Stefan Hechl

Digital interfaces of historical newspapers: opportunities, restrictions and recommendations

Eva Pfanzelter[1]*, Sarah Oberbichler[2], Jani Marjanen[3], Pierre-Carl Langlais[4], Stefan Hechl[5]

[1,2,5]University of Innsbruck, Austria
[3]Universtity of Helsinki, Finland
[4]University Paul-Valéry Montpellier, France

*Corresponding author: Eva Pfanzelter (Eva.Pfanzelter@uibk.ac.at)

Abstract
Many libraries offer free access to digitized, historical newspapers via user interfaces. After an initial period of search and filter options as the only features, the availability of more advanced tools and users' desire for more options ushers in a period of interface improvement. However, this raises a number of open questions and challanges. For example, how can we provide interfaces for different user groups? What tools should be available on interfaces and how can we avoid too much complexity? What tools are helpful and how can we improve usability? This paper will not provide definite answers to these questions, but it gives an insight in the difficulties, challenges and risks of using interfaces to investigate historical newspapers. More importantly, it will give ideas and recommendations for the improvement of user interfaces and digital tools.

keywords
historical newspapers; interfaces; digital newspapers; topic modeling; frequencies

INTRODUCTION

Interfaces for digitized, historical newspapers offer great opportunities for many different user groups, such as academic researchers, lay historians, students, teachers, etc. Interfaces steer what users can learn from digitized newspapers and they influence workflows by offering functions and tools [cf. Jarlbrink and Snickars 2017]. On the other hand, users are often not aware of biases in the search results which come from the processing and datafication of newspapers. To show the oppurtunites and pitfalls of interfaces and tools on examples from research practices is the main aim of this paper. Therefore, it focusses on an intensive testing of interfaces and digital tools. Researchers of the DH team[1] from Austria, Finland and France of the EU-funded research project "NewsEye: A Digital Investigator for Historical Newspapers"[2] experimented extensively with existing newspaper search interfaces of the three national libraries, using specific case studies in order to be able to make plausible assertions.

[1] The following researchers contributed to this essay: University of Innsbruck: Eva Pfanzelter, Sarah Oberbichler, Barbara Klaus and Stefan Hechl; Universtity of Helsinki: Jani Marjanen and Mikko Tolonen; University Paul-Valéry Montpellier: Marie-Eve Thérenty, Pierre-Carl Langlais and Nejma Omari; University of Vienna: Martin Gasteiner.

[2] This work has been supported by the European Union's Horizon 2020 research and innovation programme under grant 770299 (NewsEye).

The interfaces used were ANNO (ONB – Austrian National Library), Digi and Korp (National Library of Finland), as well as Gallica and Retronews (BNF – National Library of France).
In addition, the involved DH teams set out to use various tools that are publicly available online to analyze the newspaper corpora outside the interface environments. In order to be able to do this, they had to create subcorpora of the data they had been working on since the extraction and analysis of larger datasets from the national libraries at this point proved to be an impossible task. The creation of subcorpora therfore turned out to be one major and tricky first step in newspaper research. Using these subcorpora, they then explored different ways of performing frequency analyses, topic modelling and context analysis.
[Ehrmann, Bunout and Düring, 2019] published a survey on 24 interfaces for digitized historical newspapers with the aim of mapping the current state of the art and identifying recent trends with regard to content presentation, enrichment and user interaction. In this survey, they confirmed the still existing gap between growing user expectations and current interface capacities. While most interfaces allow users to search, filter and view newspapers, there is still a lack of content management and enrichment by NLP methods. This conclusion also applies to the three interfaces examined in this paper. However, on top of these results, the NewsEye DH team tried to find out how far they can advance their own research with the existing interfaces and where problems, distortions and limitations arise.
As an overall result, it can be stated that the NewsEye DH team was able to draw some important conclusions. Concisely said: Without the possibilities to create subcorpora, humanities researchers cannot make valid assumptions beyond studies of conceptual history. Keywords are still key and frequencies should be state of the art for contextualization and content analysis not only for academic users. Finally, topic modelling holds many promises for future development. So far, OCR quality, faulty images and metadata as well as missing functions for analysis and visualization complicate and restrict the use of the newspaper interfaces. On the other hand, existing tools outside the interfaces lack transparency, user friendliness and the possibility to process large corpora of messy data. In this sense, although some interesting hypotheses could be generated and some results could be used in support of existing studies, there is ample opportunity for improvement, if newspaper interfaces should be used as reliable gateways to in-debth analysis of these extensive archives of cultural heritage.

I SURVEY ON DIGITAL NEWSPAPER USAGE

As a part of the wider effort to examine the interest in, and use of, digitized newspaper collections across Europe, as well as to find out how the general public interacts with them, the NewsEye Digital Humanities team together with the national libraries of Austria, Finland and France in 2018 and 2019 conducted user surveys on the different online interfaces offered by the libraries. All in all, the survey showed an acceptable level of happiness with the basic functionalities of the newspaper interfaces. It appears that the user interfaces offer interesting tools, but not everyone finds them easy to use. Advanced search features, for example, sometimes require additional explanations and are hard to understand especially for unexperienced users. It also became clear that the advanced keyword search preferred by experienced users is not as well accepted as expected outside academia. On the other hand, respondents have made extensive use of filter functions, and they are also very keen on the download options. Interestingly enough, many respondents are aware of limitations of the interfaces and have many suggestions for additional tools that could be implemented. While some users do not seem to know some of the vocabulary used in the survey (OCR, metadata, full text search, etc.), they were bothered by (OCR) errors in their findings. Interestingly enough, although respondents would like many features to be implemented in the search pages, they are not willing to pay for such a service.

The wishes of users were very much in line with the experiences of the NewsEye team, which in May and June 2019 set out on an extensive testing of the mentioned interfaces. The overall results of these sessions are condensed in the following paragraphs.

II INTERFACES, METHODS, AND TOOLS

In order to be able to compare results, the NewsEye DH team first focussed on thoroughly using the existing newspaper interfaces in order to further their insight into some of the historical case studies they are conducting and, secondly, used tools and methods openly available on the internet. The starting point, however, were the newspaper interfaces.
The inerfaces that were the subject of the investigation are part of the Austrian, Finnish and French national libraries. The platform ANNO is a project run by the Austrian National Library. Accesible via the link http://anno.onb.ac.at/, it offers free and open access to 21 million Austrian newspaper and magazine pages, which were published between 1568 and 1948. Finnish digitised newspapers are searchable through two interfaces, Digi (https://digi.kansalliskirjasto.fi/etusivu), provided by the the National Library of Finland, and Korp (http://korp.csc.fi), provided by the Finnish Language Bank. Both the National Library and the Language Bank provide data dumps of the newspaper material that are openly available up to the year 1919. Digitised French newspapers can be consulted via two interfaces: Gallica (https://gallica.bnf.fr/) as well as Retronews (https://www.retronews.fr/), a platform exclusively dedicated to the historical press.

2.1 Case studies on migration, nationalism and gender

The DH researchers are working on three different case studies, which were used as a basis for the testing of the tools and the interfaces.
At the University of Innsbruck, the topic of return migration is researched with the help of historical newspapers. This case study attempts to reveal return migration processes to Europe between 1850 and 1950 asking the following questions: How and in what context were Austrian daily newspapers reporting on returnees and how did the reporting change over time? As we know today, a big number of the people who had left their homeland voluntarily or involuntarily returned to their country of origin. Austrian daily newspapers regularly reported on the return of migrants, which is why this medium is a suitable source for research on the hitherto neglected topic of return migration.
At the University of Helsinki, DH researcher focus on the topic nationalism trough newspapers. This case study uses newspapers as a text dataset to statistically describe how the language relating to nationhood changed and how newspapers themselves contributed to shaping national discourses with ethnic, class-based and gendered tensions. They use methods from natural language processing to describe how terms like nation, national, nationalism (and other -isms) changed in meaning over time.
At the University Paul-Valéry Montpellier, the topics ‘women's suffrage’, ‘women writers’ and ‘women and pants’ are investigated. Gender was an important theme across Europe during the period studied by NewsEye (1850–1940). In many ways, newspapers strengthen and spread discrimination and sexual prejudices experienced by women but at the same time, the press plays an important role in the progressive transgression of certain barriers. Some women managed to introduce literary and journalistic circles while others used newspapers to fight for the right to vote or to dress as they want.

2.2 Creating a subcorpus

Considering the amount of hits usually received when performing keyword searches on the newspaper interfaces, it quickly became clear that it is imperative for researchers to be able to create subcorpora of the available datasets. Without limiting the number of results for further and more detailed analysis, humanities researchers would often not be able to analyse the content or use it for their qualitative research questions appropriately. None of the tested interfaces offer tools to select, structure, or build datasets. The DH researchers were forced to create their own collections manually, which was extremely time-consuming.

2.2.1 Subcorpus on return migration (1864–1944)

The empirical work on this kind of research questions must be based on suitable subcorpora. For example, topic modelling (the framework of statistics-based algorithms used to identify and measure topics within a corpus) on the issue of remigration using entire newspaper editions would yield few results, since return migration was not talked about widely, especially if put in relation with the whole corpus. Also, questions regarding word frequencies within a particular topic cannot be answered easily when the entire corpus is being used. Even though distance searches allow users to search for terms appearing together in a specific surrounding, the words might still appear in different contexts. Only article separation (every article recognised as a single unit) and the possibility to address frequencies by article could prevent hits that belong to two different articles to be put together in one result. This could ultimately lead to better results when researching specific topics in the entire corpus. On the other hand, to create specific subcorpora from a large collection, both knowledge about the historical background, the research area, and about the factors of changing language over longer time periods is required. In order to create a corpus that can deliver valid research results, all of these factors are important.
For the case study on return migration, a subcorpus from the ANNO dataset is useful, as questions about types of returnees, nationalities, countries, and themes can be better answered by separating articles on return migration from the rest of the dataset. Also, qualitative analysis can be performed better on a selected subcorpus. The subcorpus on return migration contains 472 articles on remigration issues from 1864 to 1944 with 109 031 total words and 22 193 unique word forms, divided into 77 documents (each document covering one year). The subcorpus is based on articles from two newspapers available in ANNO: *Neue Freie Presse* (1864–1944) and *Illustrierte Kronen Zeitung* (1908–1944). The articles were copied from the data versions and saved as TXT, DOC, and CSV files. In the selection process, the relevance of the articles for the topic and the OCR quality were taken into account. Some errors were corrected manually, and many line breaks were removed. Overall, however, the original text quality was maintained. The following keywords were used to create the corpus:

keywords	keywords
'heimat zurückkehren'~20 ('returning home')	*'heimat rückkehr'*~20 ('return home')
'Heimgekehrten' ('returnees')	'Heimgekehrten' ('returnees')
'emigranten rückkehr'~20 ('emigrants return')	*'Heimkehrer'* ('returnees')
'Rückwanderer' ('returnees')	

Table 1. Keywords used for the subcorpus on return migration

The 472 selected articles deal with different kinds of return migration. Even though the corpus is only an excerpt of the reports on the topic in ANNO, the selected articles have been chosen to cover all types of reporting and content in order to allow different research questions to be answered. However, following challanges arose during the selection process:

- Keywords: Especially the keyword search '*heimat zurückkehren*'~20 ('return home') and '*heimat rückkehr* '~20 ('returning home') led to many results that were not related to return migration. Many newspapers and articles had to be excluded manually. This is possible if a smaller subcorpus is created, but difficult with very extensive corpora.
- Time: In order to create a corpus containing nearly 500 articles, almost 50 hours had to be invested. This process could be much faster and more efficient if a function allowed the exclusion of articles that have nothing to do with the research topic. An adequate topic modelling tool, for example, could be a useful instrument in this case. It could help to identify the context around the hits and classify the articles into relevant or irrelevant articles for a specific research question.
- Workspace: In addition, a personal workspace would also greatly facilitate the work and help to save, exclude, organise or visualise the articles. In order to create a frequency analysis, for example, Microsoft Excel had to be used.

Figure 1 shows the absolute frequencies of the subcorpus on return migration. While the *Neue Freie Presse* was published throughout the period studied, the *Illustrierte Kronen Zeitung* was printed between 1908 and 1944 only. Due to OCR problems, hardly any articles could be found in the *Neue Freie Presse* after 1930. In the *Illustrierte Kronen Zeitung*, text recognition was consistently good.

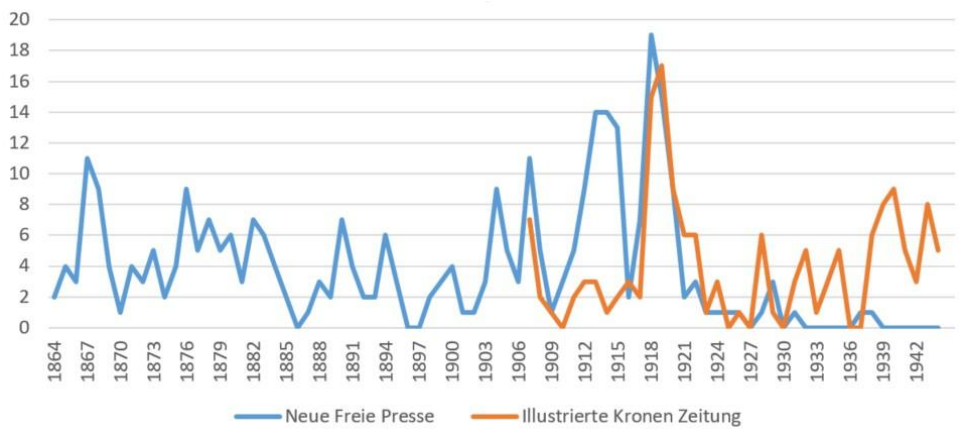

Figure 1. Absolute frequencies of the subcorpus on return migration

Since the subcorpus was created in a Microsoft Word document, metadata is missing. While information about the newspapers itself (title, publication year, etc.) can be found in ANNO, details about the content (number of words, pages, languages etc. per year/month/day, document length) are not available.

2.2.2 Subcorpus on woman suffrage (1890–1940)

The subcorpus on woman suffrage contains 628 articles from *Le Matin* and *Le Petit Parisien* published from 1890 to 1940 and available in the BNF newspaper dataset. The corpus includes the articles containing the French expression '*vote des femmes*' ('women's suffrage') in these two daily newspapers (figure 2).

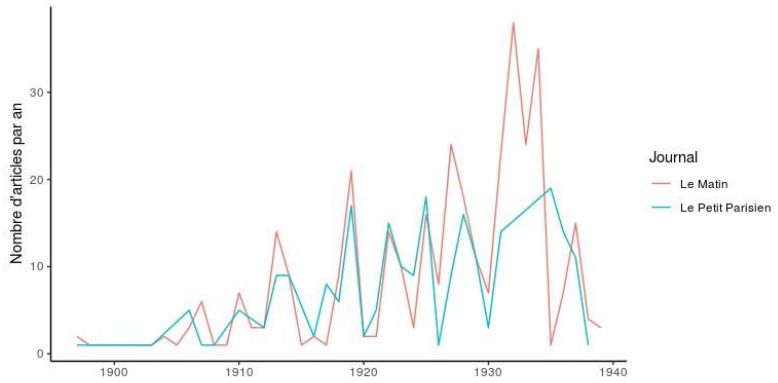

Figure 2: The articles of the French women suffrage corpus per year in *Le Matin* and *Le Petit Parisien.*

It was possible to extract articles thanks to the Optical Layout Segmentation done by the Europeana Newspaper projects in 2011–2014: Paragraphs and text blocks were grouped into articles. This digital reconstruction of articles is not perfect and is especially faulty for advertisements, which tend to be printed in messy sections, regardless of the actual segmentation between ads. Nevertheless, standard articles with a title are recognised correctly enough in order to be suitable for research purposes.

The complete digital archives of the two newspapers were already available within the programming interface of the *Numapresse* Project, hosted on the French national digital infrastructure for the humanities and social sciences, *Huma-Num*. The original dataset includes numerous supplementary pieces of metadata at the word level, for instance, the coordinates, the size or the font style for each token. This contextual approach to text mining has been described by [Langlais, 2019]. Here, only the aggregated raw text for each articles published in *Le Matin* and *Le Parisien* from 1890 to 1940 was used and the 628 publications with the string '*vote des femmes*' retained. This selection is only a small portion of available material and shows a tiny sample of the journalistic debate on women's suffrage. To get a wider perspective, it would have been necessary to use many alternative expressions (e.g. '*suffrage féminin*') or neighbouring searches (such as targeting all the articles where the word '*femme*' is close to the word '*politique*').

The quality of text recognition can affect the results as well. OCR accuracy is usually set between 80–90% for this period but can go as high as 95–99% during the 1930s, probably thanks to an enhanced quality of the original news archive. This discrepancy of text recognition across the period can potentially skew the results: We may have more articles from the 1930s simply because the query matches more results.

As the original purpose of this part of the project is to experiment with several tools and not to produce a general inquiry in a social science perspective, a smaller corpus seemed appropriate.

2.2.3 Subcorpus on nationalism

The subtopic on nationalism focuses on the long-term changes in the vocabulary relation to nation in Finnish newspapers, with a starting point in 1771 and ending in the 1920s. Simple free-text searches yield up to two hundred thousand hits for the words '*nation'* in Swedish and '*kansakunta'* in Finnish. Hand curating a subcorpus of articles based on the search results would be possible only if the selection was to be narrowed down remarkably to a particular aspect of writing about nation, like in some of the previous subtopics. The path chosen for this subtopic was rather to choose a selection of empirical cases that were studied qualitatively, but then doing the bulk of the analysis by using descriptive statistics with all the instances

containing words belonging to the vocabulary of nation. Some of the results are described below under the heading 'Frequency analysis'.
However, for some of the analyses produced for the subtopic, subcoprus creation was in effect done by using search terms and downloading the Key Words in Context (KWIC). Contexts were created either according to a set window (five or more words before and after the selected term) or according to a flexible context (downloading whole paragraphs or sentences in which a particular word appeared in which cases the length of the window varies). These subcorpora were used in particular when producing word-vector-space models for analysing in which contexts the vocabulary of national was used in Finnish newspapers. They were also used in producing frequency lists of the lexical context of nation, national and nationalism.
The method of focusing on a set window or flexible context is not as accurate as using identified articles as a context for analysis. As the Finnish newspaper dataset does not include reliable article segmentation, this, however, remains our best method of doing contextual analysis without hand-curating the subcorpus. Using a set window has the benefit of producing symmetrical contexts for each keyword search, but this method is ignorant of sentence structure or stylistic features in the text. If the window is large, say one hundred words before and after, it is also more likely that the chosen context includes text from articles that precede or succeed the target article. Focusing on a sentence or a paragraph has the benefit of rarely transgressing article boundaries, but this method is prone to get the context wrong due to OCR-errors regarding punctuation. Further, focusing on the sentence or paragraph as context gives more prominence to articles with long sentences or paragraphs. For these reasons the set window worked with is fairly small.

2.3 Frequency analysis

For historians working with large quantities of digitised texts from long periods of time, the possibility to not only note the introduction of new vocabulary, but also to be able to quantify the frequency of words or other features of language has become a way of illustrating changes in how people conceptualised the world. Newspapers form an especially interesting data set for this, as they are, relatively speaking, a rather stable corpus that deals with a large selection of topics. The first step when working with an adequate corpus using digital tools therefore usually is to perform a simple frequency analysis. This can give researchers a quick overview of word frequencies (or at least frequencies of combinations of letters) and trends in the selected dataset. In addition to co-occurrences or topic modelling tools, frequency analysis can be a first entry point for more complex investigations of larger and complex datasets.
It is common practice in humanities scholarship to trace the first occurrence of a word relating to the topic of a study. As [James and Steger, 2014] discuss, this is sometimes called 'the religion of the first occurrence'. Digitised sources are obviously helpful in identifying new candidates for first uses of particular terms through keyword searches, but more importantly, they make it possible to shift focus from early uses to the analysis of frequency and answering questions of when it became more common in general to use a particular word or when only some people chose to use a particular terminology.
Digitised newspapers are a very good dataset for diachronic analysis of word frequency for many research questions, as it is relatively easy to select subcorpora based on particular newspapers (perhaps with different political leanings), particular places of publication or the publication frequency of newspapers. To compare changes in word frequency over time, frequencies need to be normalised to account for different sized corpora for each time slice

(usually a year), so that we get a relative frequency of words used per one million (for instance) words in the corpus of each year.
For corpus linguistics, the analysis of word frequencies is much about getting to the truth of language use and hence having a balanced corpus is imperative. For historical purposes, more uncertainty is acceptable and also unavoidable because of the nature of historical research questions. The historians' interpretation about what changes in word frequencies are about must also try to account for changes or biases in the data. In general, changes in word frequencies are about:

- a word becoming very topical in a given moment
- a word entering new domains in language
- a word being associated with new meaning (polysemy)
- something data-specific that skews frequency. In the case of newspapers this might for instance be repeated advertisements that are surely part of the historical dataset, but would be cleaned out if the data set was seen as balanced in a corpus-linguistic sense.

2.3.1 Frequency analysis on nationalism using Digi and Korb

The subtopic on nationalism relies on digitised newspapers as a massive text dataset that can be used to trace long-term changes in the vocabulary relating to nation, national, nationalism and related terms. As such, it provides an example of the possibilities and limits of using frequency analysis in the digitised Finnish newspapers to study complex historical phenomenon.
One first step is to understand that words like nation and nationalism should not be studied alone, but in the context of related terms, but at the same time it is clear that individual words do have their own trajectories and that choosing to talk about nationalism and not just nation has most probably been a conscious choice by a historical actor. Typically, the first uses of the words 'nation' and 'nationalism' have been mapped thoroughly, for instance by [Kemiläinen, 1964]. One of the most famous and early uses of the word 'nationalism' is by [Johann Gottfried Herder, 1774] in *Auch Eine Philosophie der Geschichte zur Bildung der Menscheit* (in English 'This Too a Philosophy of History for the Formation of Mankind'). Herder writes about the animosity between nations and how Herder's philosophical/political opponents label this as 'prejudice! mob-thinking! limited nationalism!' As Herder is one of the main figures in the history of national thought, this use of nationalism has naturally been noted, but from the point of frequency of the use of the word 'nationalism' e.g. in Swedish-language newspapers, the issues come across differently. Early uses, such as the one by Herder, appear as being occasional uses, whereas 'nationalism' becomes more frequently used only at the end of the nineteenth century and beginning of the twentieth century (figure 3).

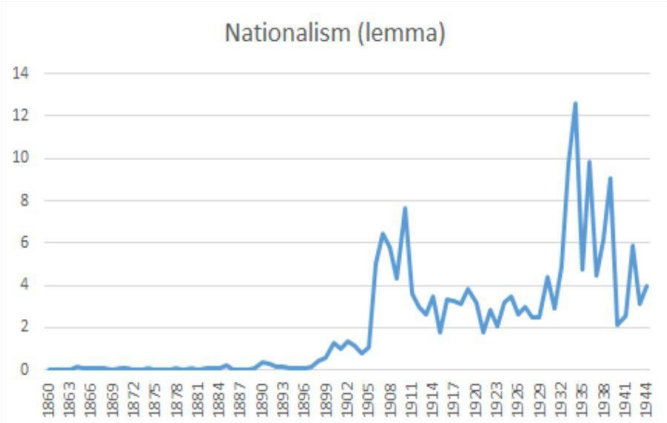

Figure 3. Relative frequency (hits per one million words) of the lemma ‘nationalism’ in Swedish- language newspapers in Finland. False positives resulting from bad OCR of the word ‘rationalism’ as ‘nationalism’ are omitted.

A reading of some of the uses of ‘nationalism’ in the late nineteenth century shows that the word was used in a negatively laden way in heated debates about Swedish, Finnish and Russian nationalism in Finland. The rise in frequency of the term ‘nationalism’ at this time was not only a Finnish phenomenon, but can be seen also in other similar datasets, such as the manually created absolute frequency figure 4 from ANNO. The growth in frequency for ‘nationalism’ was clearly about the word becoming topical in certain debates.

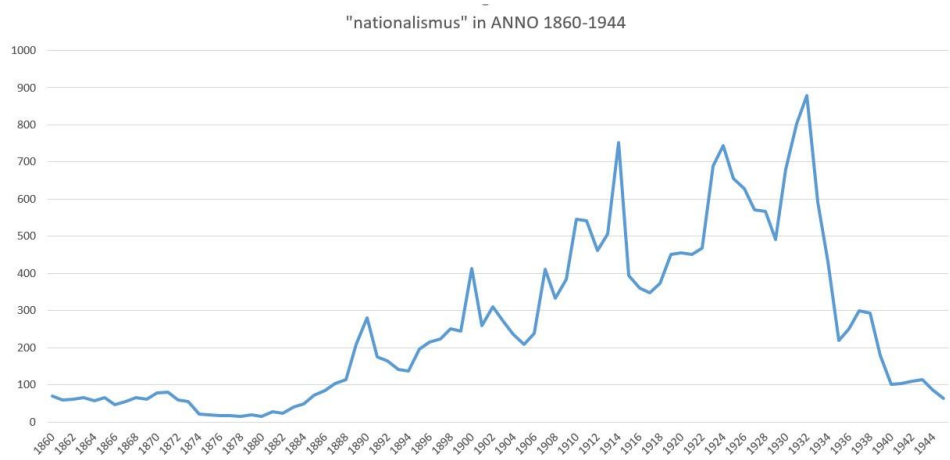

Figure 4. Absolute frequency of the keyword ‘*nationalismus*’ in German language newspapers in Austria. Early hits include many false positives due to OCR errors in reading the word ‘*rationalismus*’ like in the Finnish case. Bad OCR is also corrupting the results from 1930s onward (see below).

If we turn to the word ‘national’, the story is also a story of growth during the nineteenth century, but still slightly different (figure 5).

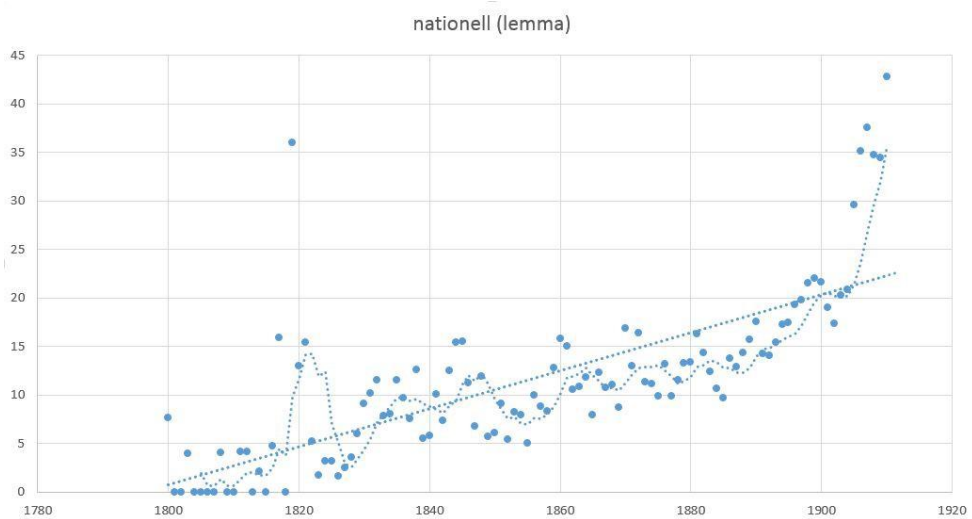

Figure 5. Relative frequency of the lemma ‘*nationell*’ in Swedish-language newspapers in Finland, 1800–1910.

The fact that the relative frequency of 'national' and 'nationalism' do not quite correspond suggests that the historical use of both terms is related, but that they address slightly different things. Hence, also the trends in their use differ. Looking at not only the frequency of 'national', but also looking at the context of the term is in this sense revealing. Since 'national' is most often followed by a noun that tells us which things could be conceptualised as being national, we can also count how many different nouns co-occur with 'national' each year in the newspaper corpus. This is called the productivity of 'national' (figure 6).

Figure 6. Productivity in the bigrams starting with the word '*nationell*' ('national') in Swedish-language newspapers and periodicals in Finland, 1800–1945. After 1910, the size of the annual data sets is smaller.

The increase in productivity indicates that over the course of the nineteenth century, the word 'national' was used in more and more domains of language. By the end of the century, the nouns co-occurring with 'national' related quite different spheres of life from 'national economy' and 'national assembly' to things like 'national anthem' or 'national language'. The proliferation of nouns co-occurring suggests that also the general growth in relative frequency is more about the national perspective becoming nearly all-encompassing. Only in the years 1820 and 1821 do we see a clear peak in relative frequency that can be explained by a factor in the data. In those years, the publications *Mnemosyne* and *Åbo Underrättelser* radicalised the language of 'national', which is also clearly visible in the data.

Frequency analysis for the purpose of historical interpretation still relies on a heavy interpretative component. Historians are often interested in changes in language, but also relating those changes to societal change, which always requires a discussion about how changes in language use relate to other changes in society. In the cases of the words 'national' and 'nationalism', we see that while the words share a root, the concrete uses of the terms differ. As pointed out by [Kurunmäki and Marjanen, 2018], this is much due to the rhetorical function that is very often accompanied with the use of different 'isms'. Studying the frequency of the word 'nationalism' is thus not a good indicator for studying how people understood the nation or even less how nation-building as a process advanced in Finland (or elsewhere), whereas looking at the frequencies of 'national' tells us about the process in which the national perspective became dominant and therefore also relates more closely to nation building as a process.

For a critical discussion of Google Books and some of the work drawing on frequency analysis in them, see [Pechenick, Danforth, and Dodds, 2015]. For historians to better understand to which extent relative frequency can be used to draw conclusions relating to

historical processes and get beyond some of the naive interpretations that have emerged especially after the Google Ngram Viewer was made public, they need better tools for using frequency measures for explorative work. As producing frequency graphs is (computationally) not a demanding task, frequency analysis should not be left to the projects that can download full data of digitised data sets and process them, but should ideally also be available as integrated tools in graphical user interfaces – as is the case, for instance, for digitised Dutch newspapers at [KB Lab].

2.3.2 Frequency analysis on return migration using ANNO

Since the literature dealing with the issue of return migration to Europe is limited, little is known about remigratory processes between 1850 and 1950. A frequency graph could therefore reveal when remigration became an important issue in the newspaper coverage and as a result an issue for the societies people returned to.
ANNO does not yet provide any tools to create frequency analyses automatically. Nevertheless, with the help of the information provided by ANNO (absolute number of search terms in a certain period of time) it is possible to create frequency graphs by manually copying the numbers to a spreadsheet outside of the platform. The absolute numbers of hits per year are displayed in the filter section of ANNO and can be extracted manually (figure 7).

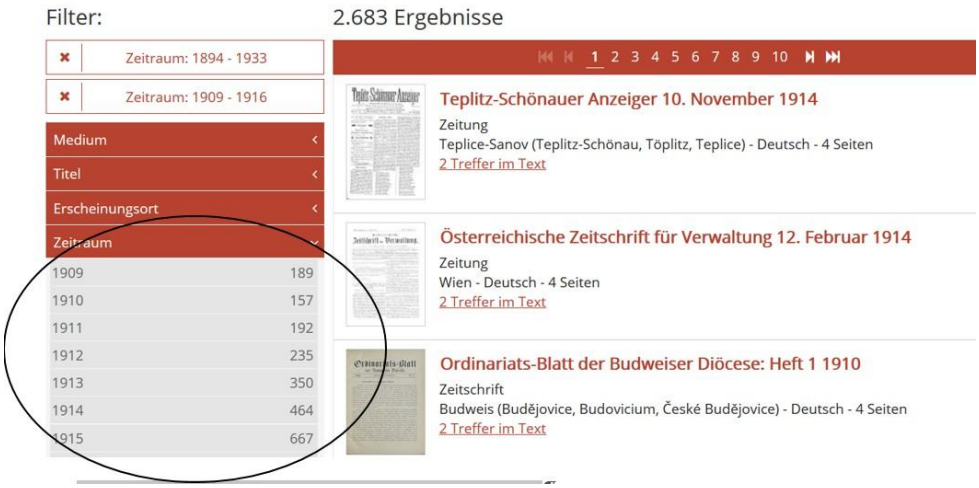

Figure 7. Hits of the search term '*Heimat zurückkehren*'~20 in ANNO.

After entering the obtained hits in an Excel file, an absolute frequency graph with the search terms '*Heimat*' in combination with variations of '*rückkehr /zurückkehren/heimkehren/zurückgekehrt*' ('home' combined with 'returning/return/return home/returned') could be created (figure 8). Since the distance search function does not allow terms to be marked with an asterisk (*), many term combinations were required. For example, it is not possible to search for '*heimat *rückkehr**'~20 ('home *return*').

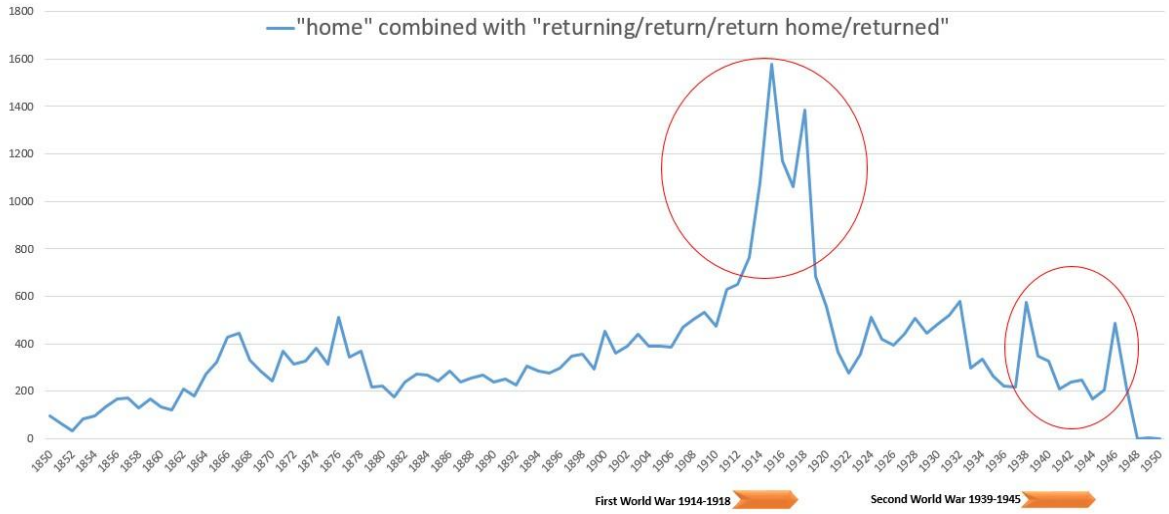

Figure 8. Frequency graph of 'Home' combined with 'returning/return/return home/returned' (1850–1950).

Taking historical events into account, a closer look at the graph immediately revealed a discrepancy: The number of hits, not unexpectedly, rises sharply during and at the end of the First World War, but very surprisingly not during the Second World War. This discrepancy also became apparent with many other search terms. The question of why the results of the frequency analyses were distorted led to a consideration and subsequent examination of the following questions:

- Were important search terms overlooked?
- Is there a discrepancy between the number of published newspapers during the First World War and the Second World War?
- Is there a discrepancy between the number of digitised newspapers during the First World War and the Second World War?
- Is there a change in the quality of the digitised newspapers or the OCR for certain newspapers?

In order to facilitate the interpretation of the results, the same keyword search ('home' in combination with 'returning/return/return home/returned') was carried out by selecting only one newspaper. This made it easier to estimate the quality of the OCR, the number of issues published, and the volume of the dataset per newspaper. The newspaper *Illustrierte Kronen Zeitung* (1905–1944) seemed to be suitable for this, as the quantity and the OCR quality remained more or less the same throughout the entire time period. As it turned out, however, there is no OCR text available for the years 1936 and 1937. This limited search delivered different results. In contrast to figure 8, figure 9 shows a clear increase in the usage of the search terms at least for the beginning of the Second World War:

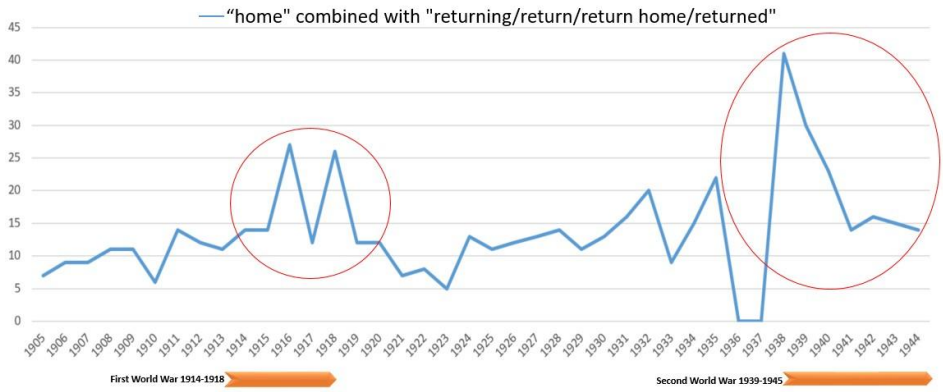

Figure 9. Frequency graph of the search term 'Home' combined with 'returning/return/return home/returned' in the newspaper *Illustrierte Kronen Zeitung.*

Therefore, figure 9 shows clearly that missing keywords are not the main problem here. Instead, two issues were identified:
First, absolute frequencies can show an unrealistic picture, as they only count the number of times a word (or a combination of letters) appears in a corpus without putting the results in relation, e.g. with the total number of words for a certain time period or a certain newspaper. However, relative frequencies cannot be generated with ANNO data as it exists now, as required metadata is not available, i.e. word count or number of digitised pages/newspapers per year.
Second, very time-consuming browsing and comparing of the two selected newspapers and their OCR'd versions revealed that the quality of the OCR between 1938 and 1945 varies extremely from newspaper to newspaper. Figure 10, for example, shows an article on return migration in the newspaper *Neue Freie Presse* from 1938. On the left, readers can see the scan of the original text and easily read it. On the right, the OCR'd letters do not resemble German words at all. Apart from the headline, almost no word is readable and many of the characters shown are not part of a normal German text.

Rückkehr in die befreite Heimat

Donnerstag mittag trafen in Reichenberg die ersten in die Heimat zurückgekehrten Flüchtlinge ein. Es handelt sich um 16 Männer aus dem Bereich der SA-Gruppe Mitte (Magdeburg), die während ihres Aufenthaltes unter dem Schutze des Altreiches von der Gruppe SA-gemäß ausgebildet, vollkommen eingekleidet und jetzt in ihre Heimat zurückgeleitet wurden. Zum Empfang versammelte sich die gesamte Bevölkerung auf dem Marktplatz, wo gegen Mittag sturmweise die Männer mit gepacktem Tornister, voran die Fahne, auf dem Adolf-Hitler-Platz aufmarschierten. Besonders stürmisch begrüßt von den Heimkehrern wie von den Reichenbergern wurde bei seinem Erscheinen der Reichskommissar. Nach der Begrüßung durch einen Vertreter der Stadt und nach einer Ansprache von SA-Obergruppenführer Kob, erklärte Konrad Henlein: Alle, die hier standen, seien bereit gewesen, ihr Leben für die Heimat in die Schanze zu schlagen. Daß sie nicht eingesetzt zu werden brauchten, dankten sie dem Führer. Konrad Henlein gedachte dann der Toten, die um Sudetendeutschlands Freiheit gefallen waren. Sein Dank galt den Betreuern der Flüchtlinge im Reich. Die Kameradschaft und begeisterte Aufnahme im Schutze des starken Reiches sollten die Männer durch ihre stete Einsatzbereitschaft danken. Nach dem Sieg-Heil auf den Führer zogen die Formationen unter Vorantritt des Musikkorps und des Spielmannzuges der Gruppe Mitte an Konrad Henlein und den Gruppenführern vorüber.

Rückkehr in die befreite Heimat
®onnerStag mittag trafen in IReidhcubcrg bie
e r fte n in bie .§cimat zurückgekehrten f5tkrhi^iitgo ein.
h®>^bclt sich 16 SJlänner auS bem Sereidt ber @91»
©nippe HUitte (ünagbeburg), bie währenb il)reS 9lufentholte§
unter bem ©d)uhe bcS 9lltreid)e§ oon ber ©ruppc '^»gemiä^
ausgebilbet, oollkommen eingekleibet unb je|t in ihre .^eimat
Znrückgeleitet würben. ©mpfang oerfammelte sich bie
gesamte ©eoöllterung auf bem iDlarktplah, Wo gegen Sllittag
ftnrmweife bie SRäuner mit gepadrtem Tornister, Uoran bic
f^ahnc, auf bem 9tbolf»,§itler»ißlah aufmarfdperten. ®e=
fonberS stürmisch begrüfd non ben .^eimkehrern wie oon ben
ikeidjenbergern würbe bei feinem ©rfcheinen ber fRcichiS»
kommiffar. 9ladji ber Segrü^iing burd) einen 9?ertreter bet
©tabt unb itod) einer 9tnfpta<he toon ©2l--Dbergruppenfül)rer
,t 0 b, erklärte tonrab § e n l e i n : 9ll[c, bie hier fton^n,
feien bereit gewesen, ihr Sehen für bie ^cimat in bie ©chanze
ZU fd)lagen. ®a^ sie nidjt cingefeht zu werben brausten,
bankten sie bem ü l) r e r. tonrab §entein gebachte bann
ber Soten, bie um ©iibeteiibeutfchlanbS gefallen
waren. Sein ®ank galt ben tBetreuem bet f^tüchtlinge tim
llieich. ®ie tamerabfdjaft unb begeisterte 9lufnahmc im
©cljuho beS starken i){eid)e§ sollten bie üUänncr burch ihoo
stete ®infahbereitfd)iaft banken. 9foch bem ©ieg=.g>cil ans ben
f^ührer Z'ogou bic f^onuotionen unter 35orantritt be§ SOluftk»
korpg unb be§ ©pielmaiinzugeä ber (%uppe SOlittc on tonrab
^nlcin unb ben ©ruppenführern borübet.

Figure 10. *Neue Freie Presse*, October 15th 1938 (PDF, left / TXT, right).

To summarize, it can be said that the OCR quality in ANNO varies significantly not only over time but also between different newspapers. The above results led researchers to become extremely skeptical about the results they achieved with the full text searches in the newspaper datasets. Also, it became clear that the lack of metadata and the lack of adequate tools prevents the preparation of conclusive frequency analyses, which in turn could point users to some of the flaws mentioned above. The manual creation of frequencies was time-consuming and cannot seriously be considered for ordinary research efforts. Last but not least, the irregularities found in the search results lead to a more or less devastating conclusion, namely that the automatically created search results sometimes are invalid or even false.

2.4 Analysing created subcorpora using existing topic modelling tools

For historians, computer science and digitisation have held many promises. From the early 1970s until today, however, as [Belvins, 2016] argued, for digital history the 'sunrise of methodology was still hovering just over the horizon'. Digital methods, despite all the advances in recent years, especially concerning quantitative approaches, seem to be an eternal promise. One of those promises certainly is the potential possibility to instruct machines to automatically extract meaningful topics from huge corpora of data. It has sparked the discipline's interest in the computer scientists' topic modelling approaches ever since the introduction of the concept in the late 1990s, as described by [Papadimitirou et al., 1998].
Topic modelling tools can basically be understood as algorithms that extract meaningful topics from text. According to a tweet by [Garfinkel, 2012], a topic can be defined as 'a recurring pattern of co-occuring words'. Or, in the words of [Graham, Weingart, and Milligan, 2012], a so-called topic 'consist of a cluster of words that frequently occur together'. In a good topic model, the results – the topics or rather the words in the topic – make sense to the reader. [Brett, 2013], for instance, uses the topics 'tobacco, farm, crops' and 'navy, ship, captain' as examples.
Newspapers have been a popular subject for topic modelling, because they provide a way to study the change of words over time from a daily source, as [Brett, 2013] argues. Current and historical newspapers have been investigated with topic modelling, see for instance the work by [Nelson] on *The Dispatch*.
Since there are no topic modelling features available on the newspaper interfaces investigated here, the NewsEye DH teams used several other available tools that allow users to experiment with topic models.

2.4.1 Creating topics on return migration using STM - an R package for the structural topic model

STM is an R extension that has to be used within a program. It is centered on non-supervised text classification and does not focus on corpus preparation. STM is fairly new and documentation is a bit lacking: There is little explanation on the package page and for the current case, the well-commented use by [Silge] was mostly relied on.
STM was used with the corpus on return migration (472 articles). The corpus were prepared using naive lemmatisation with lookup tables: This is not a significant issue for text classification based on thousands of individual tokens. Stop words were not removed manually. The following 8 topics were generated for the whole corpus (figure 11):

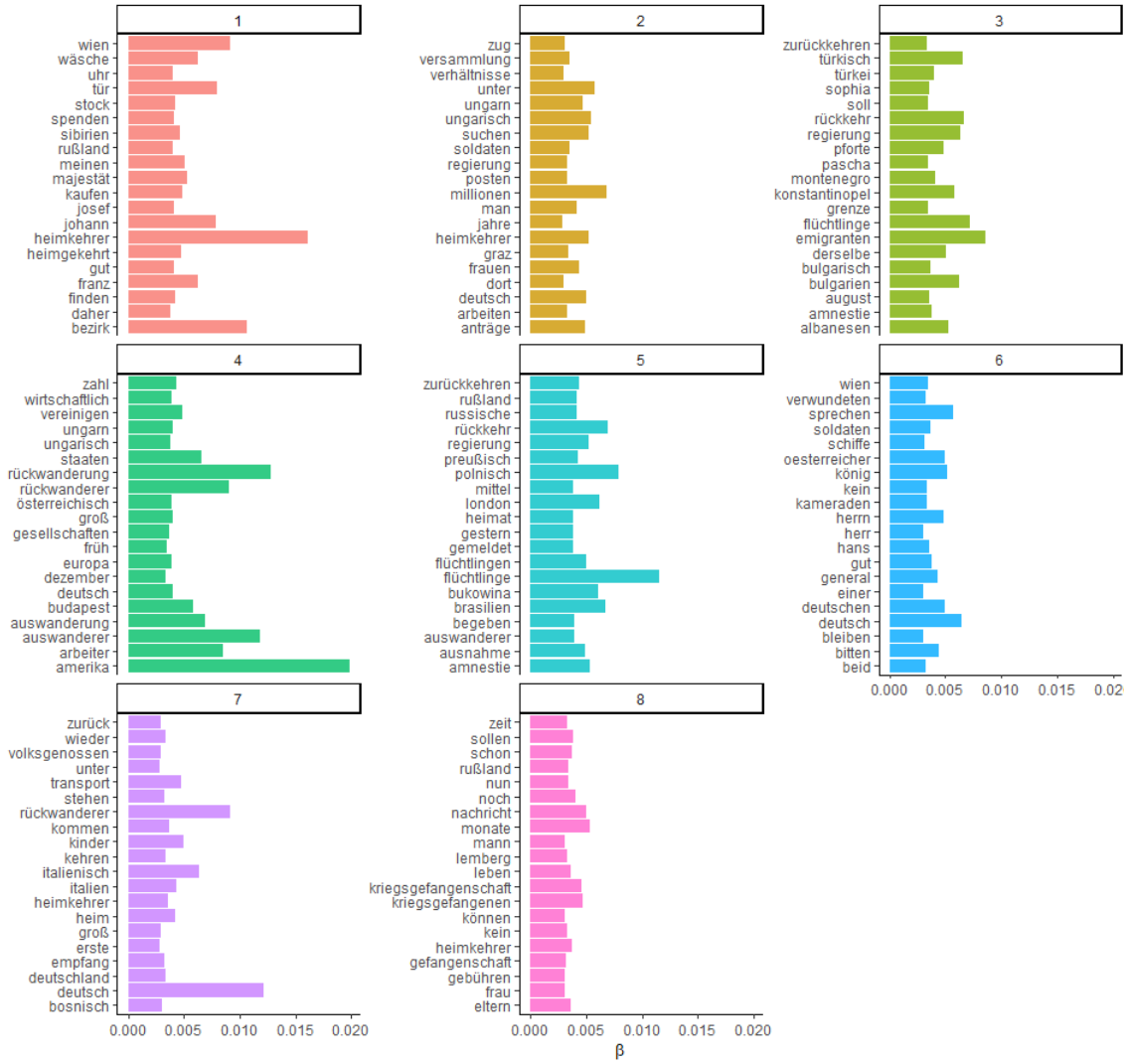

Figure 11. Topic model with STM from the subcorpus on return migration.

At a first glance and without knowing the content of the corpus, it is difficult to tell what the topics really stand for. Only topic four seems to be very clear: the return of people from America. Topic 1 seems to map the discourse of laundry donations for returnees and topic 8 seems to reflect the discourse on prisoners of war. In order to provide more accurate interpretations, a better insight into the corpus is needed.

If, for example, a closer look at topic three is taken, the question arises how the words '*türkisch/Türkei*' ('Turkish, Turkey'), '*Sophia*' ('Sofia'), '*Montenegro*', '*bulgarisch/Bulgarien*' ('Bulgarian, Bulgaria'), '*Albanesen*' ('Albanian'), '*Konstantinopel*' ('Constantinople'), '*Flüchtlinge*' ('refugees') and '*Pascha*' are related to each other. It needs deeper historical knowledge and an insight into the corpus to recognise that the Serbo-Turkish war from 1876 to 1878 could be at the centre of this topic. At the same time, topic 3 could also depict a very different historical event that took place in 1911, 34 years later. The terms '*zurückkehren*' ('return'), '*Türkei*' ('Turkey'), '*türkisch*' ('Turkish'), '*Amnestie*' ('amnesty'), '*Montenegro*', and '*Albanesen*' ('Albanian') could refer to the return of Albanian Malissors, mountain tribes, from Montenegro to Turkey in 1911. The Young Turkish government had issued a general amnesty for the Albanian refugees. Although the terms '*Sophia*' ('Sofia'), '*Bulgarien/bulgarisch*' ('Bulgaria/Bulgarian') and '*Emigranten*' ('emigrants') are not related with the return of the Malissors, they appear simultaneously (1911/12) and refer to the return of Bulgarian emigrants from America to Bulgaria.

Topic 5, by contrast, is a good example for a misleading topic. Since '*polnisch*' ('Polish') and '*Flüchtlinge*' ('refugees') are the highest rated terms, the topic seems to be about Polish refugees. But a closer look into the corpus has led to the following findings: It can be said that '*Rußland/russische*' ('Russia, Russian'), '*polnisch*' ('Polish'), '*Rückkehr* ' ('return'), and '*Auswanderer* ' ('emigrants') have something in common. The terms seem to refer to the return of Polish emigrants from Brasilia to Russian Poland in 1891. Also, the terms '*London*' and '*Amnestie*' ('amnesty') are related to 'Polish': In 1874, Polish exiles in London asked for an amnesty to return to their home country. And in 1883, an amnesty was granted to Polish insurgents with the possibility to return from Russia to Poland. In the same year, 1883, the Bukowina became an issue as well in the context of the repatriation of Csangos from the Bukowina to Austria-Hungary. The Bukowina is only insofar connected to Russia, as the region was occupied by the Russians up to 1774. Both Poland and Russia again played a role in the newspaper coverage in 1919, when Austrian prisoners of war returned from Russia, crossing Poland. In the same year, the Bukowina also again became a subject of reporting: Refugees (the most important term in topic five) from the Bukowina, who had fled to Austria, were to return home. Again, the refugees from the Bukowina have little in common with Poland or Russia (at least in the corpus), but just as in 1883, 'Polish' and 'Bukowina' appeared together. It thus becomes clear why the algorithm created this topic. Nevertheless, the Polish remigration and the return of refugees from the Bukowina are two different topics. The common ground is simply return migration during the same periods of time.

On the other hand, important topics, such as the return of Galician refugees during the Fist World War, are missing. This problem can be solved by increasing the number of topics. When generating 12 topics instead of 8, the discourse on Galician refugees appears. The question of how many topics should be generated depends on the research question and on the corpora itself.

STM delivered some impressive results. It is definitely a helpful tool to discover main topics. However, just like most of these tools, lacks the possibility to link back to the original text (that is where the topics occur in the corpus). For historians, reading the context is often the only possibility to find out why words were selected for a topic. Therefore, the interpretation of the topics remains a real challenge. The same is true if users want to make sure topics are formed from the correctly connected words. A closer look at the corpus in connection with the topics selected by STM has further shown that various topics can be interwoven. Sometimes this makes sense, but sometimes it does not. The results provided by STM must always be checked manually.

The last but most important conclusion concerns the epistemological value for the research project. Even tough the interpretation of the topics turned out to be difficult, the tool pointed out the most frequent topics in the corpus. Since topic modelling was used with the intention of finding important entry points for further (quantitative or qualitative) investigation, this research aim was achieved. A disadvantage, however, is the lack of user-friendliness. It is difficult to use the R package without any programming skills. User-friendliness should be high on the agenda of computer scientists if these tools are intended for use by Digital Humanities scholars and especially by a broader audience.

2.4.2 Creating topics on woman's suffrage using TidySupervise

TidySupervise is a new tool recently developed by the *Numapresse* project. Like STM, it is an R package. It originally started as a set of customised functions to deal with the most ambitious project of *Numapresse*: the classification of newspaper per news genre for a very long period of time (1895–1940 for both *Le Matin* and *Le Petit Parisien*).

Contrary to STM, TidySupervise relies on supervised classification. The model is built upon a manually labelled training corpus and attempts to identify the words that are regularly associated with the labels. While unsupervised models have to be interpreted at the end of the classification, here the interpretive work comes first. Supervised models consequently offer less flexibility: They take no new information into account and only apply the labels and the corpus that have already been predefined.

While supervised models require significant manual input before being ready for use, there nevertheless is a significant advantage: The results do not need to be reconstructed ex-post. The models are fixed and transferable across corpora and the processing is much quicker: It takes only minutes to classify one entire month of newspaper issues, and several hours to deal with decades. Unsupervised topic modelling on such a scale would usually take days. Recent studies in Cultural Analytics, e.g. by [Underwood, 2016], have also shown that supervised models can contribute to analyse complex cultural phenomena, including the process of genericity or hybridation accross genres, through the use of combined probabilities per texts.

For this case study, the 1920–1940 newspaper model of *Numapresse* was reused and trained on 25 issues of four dailies published from 1920 to 1940. Results may be slightly anachronistic for the articles published before 1920 (for instance by stating that some early 1900s texts belong to the movie section, which actually only appeared in 1913 in the French press).

Supervised genre classification with TidySupervise (figure 12) turned out to complement STM-reconstructed topics very well. Some cross-classifications fully converge, such as policy processes with '*institutions politiques*' ('political institutions'), the genre recording the inner workings of parliament and similar institutions (figure 13).

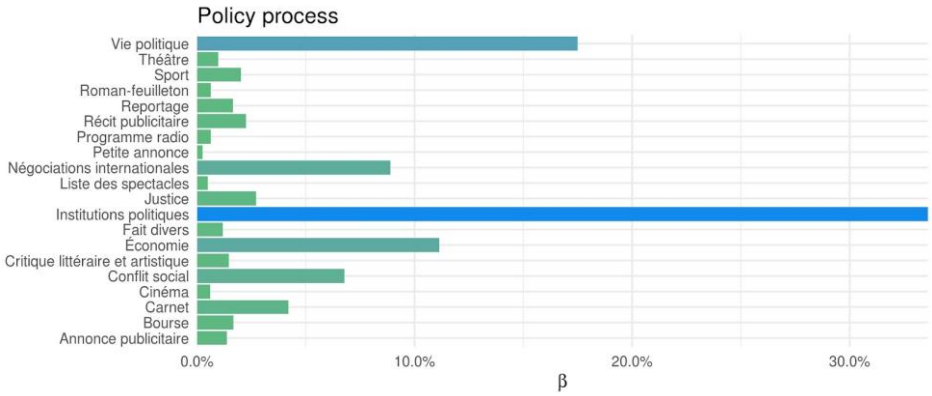

Figure 12. The supervised classifications from TidySupervise that match best the unsupervised cluster of policy processes in STM.

Other alignments are not perfect, but fairly understandable. The feminist movement is both linked to the political news ('*vie politique*' and '*institutions politique*'), but also to social conflict news ('*conflit social*'), as figure 13 shows. This actually echoes the main feature of the movement, which pushed forward a political agenda to emancipate half of the population from gender discrimination.

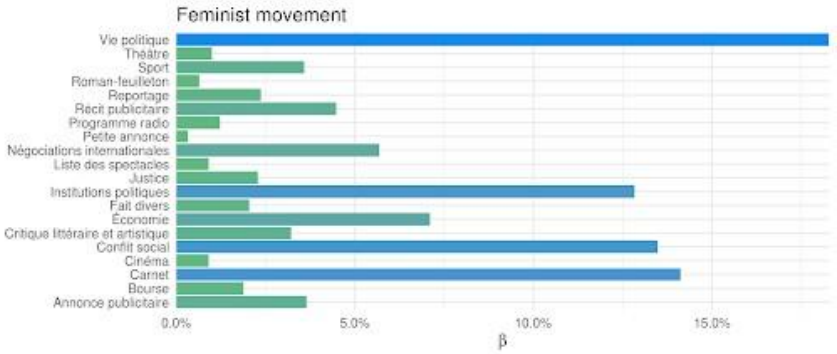

Figure 13. The supervised classifications from tidysupervise that best match the unsupervised cluster of the feminist movement in STM.

Finally, in figure 14, advertisements turn out to be frequently classified as 'announcements', hence the spread of the '*carnet* ' (register) genre. This is not really a wrong classification, since notices of feminist meetings are not necessarily paid ads. It may also be partly a consequences of limitations of Optical Layout Segmentation within the Europeana Newspaper project. Ads and announcements are generally regrouped in wider sections and there is way to identify the precise ads using the expression '*vote des femmes*'.

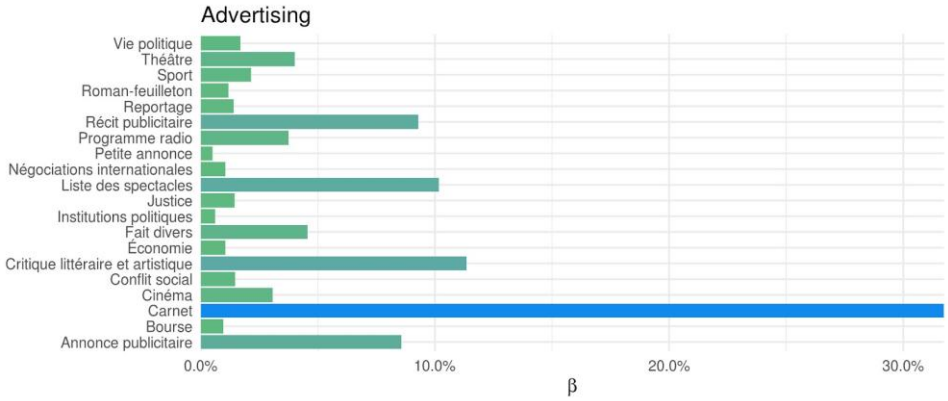

Figure 14. The supervised classifications from tidysupervise that match best the unsupervised cluster of the feminist movement in connection with Advertising in STM.

Supervised classification can also be useful to detect hapax legomena or anomalies, while unsupervised topics modelling is more focused on wider classification. In the results there are a handful of articles classified with a high probability as serial novels. This instance could still be an important signal of wider acculturation processes of the issue of women's suffrage, beyond the political column. Such a hapax would never have been singled out using only an unsupervised model, as classifications are usually derived from larger clusters of documents. Since it is drawn from an external corpus, supervised classification can highlight such unusual ouputs.

2.3.1 Creating topics on return migration using Overview

[Overview] is an open-source platform that promises to be able to automatically analyse thousands of documents. It includes full text search (including fuzzy search), visualisations, entity detection, topic clustering, etc. When it comes to topic modelling, the platform explores word patterns using a rather different process than other tools. The focus lies on the comparison of two documents in order to see their similarity. This comparison is made by

multiplying the frequencies of equal words and then adding up the results. Documents with high similarities are grouped together using a clustering algorithm based on this similarity of scores. Overview then categorises the documents into folders, sub-folders, sub-sub-folders and so on (figure 15):

Figure 15. Folders created with Overview.

Overview was used to find topics in the subcorpus on return migration. For this purpose, 77 single Word documents (each document covering one year) were uploaded to the open-source platform. When uploading, users can choose whether each file or each page should count as a document. Since not every article already represented one file, it seemed more appropriate to select the second option and prepare the corpus accordingly. As a result, each article was then referred to as one document. This was important, because otherwise, not single articles but entire files containing many articles talking about different things form the basis for the evaluation, whereby topics are mixed together. The program removes stopwords automatically (German can be chosen as a language). In addition, it is also possible to remove irrelevant words manually (like recurring titles of the newspapers etc.). The program then compares every pair of documents (535 pages in total). Just like STM, Overview recognised some of the main themes, such as the Serbo-Turkish War from 1876 to 1878, or the return from America. The topic about the return form America, for example, was found in 55 articles and the folder was labelled with the following words: '*Rückwanderer* ' ('returnees'), '*Auswanderer* ' ('emigrants'), '*Arbeiter* ' ('workers'), '*Rückwanderung*' ('return migration'), '*Amerika*' ('America'), '*Auswanderung*' ('emigration'), '*Staaten*' ('States'), '*Hamburg*'.

The same folder is split into several sub-folders with even more specific topics (each folder splits into smaller and smaller sub-folders, each of which contains a smaller number of documents). One of the sub-folders, for example, was labelled with the words '*Amerika*' ('America'), '*Rückwanderer*' ('returnee'), '*Rückwanderung*' ('return migration'), '*Hamburg*', '*Krise*' ('crisis'), '*Dezember* ' ('December'), '*allerhöchste*' ('at the very most'), '*Bremen*', '*Auswanderer* ' ('emigrants'), '*europäischen*' ('European'). This topic occurs in twelve articles and maps the news coverage in the subcorpus on remigration from America as a consequence of crises.

Because of the many sub-folders, even more hidden topics can be revealed. For example, six articles were found talking about care and help, particularly for returned emigrants. In another sub-folder, requests and small advertisements from and for returnees were correctly

put in relation to each other based on the following words: '*Heimkehrer* ' ('returnee'), '*Posten*' ('post/position'), '*verheiratet* ' ('mar- ried'), '*bittet* ' ('asks/begs'), '*Anträge*' ('requests'), '*Hans*', '*sucht* ' ('looking for'), '*bez*' (abbreviation for 'district'), '*spricht* ' '(talks)', '*gesuche*' ('requests'). Even though most of the generated topics were understandable, sometimes no obvious common topic could be recognised. In this instance it was helpful to have the original text at hand, a feature implemented in Overview.

Just like some of the other topic modelling tools tested, Overview was able to map the most important topics. Since every folder is split into smaller and smaller sub-folders, even more hidden and specific topics where found.

All in all, Overview seems to have some clear advantages over other topic modelling tools:

- The platform is very user-friendly and easy to use.
- Interactive topic modelling: The tool allows users to iteratively refine the topics by adding and excluding keywords.
- The tool displays where the topics occur in the corpus: This makes the results comprehensible and the topics understandable.
- Overview offers a personal workspace: This allows users to annotate (tag) the documents and to save the results.

III CONCLUSION

The intensive testing of the existing newspaper search interfaces ANNO, Digi and Korp, as well as Gallica and Retronews, showed both similarities and some differences, which influence the search outcomes in different ways. ANNO, for example, has the largest collection and offers a very good distance search option. Digi, on the other hand, allows users to download the entire collection for further processing and Gallica gives very useful informative historical introductions for topics that have been chosen by the designers of the interface. On the other hand, while both the National Libraries of Finland and of France have separate search interfaces for advanced analysis, the Austrian National Library only recently launched a similar page (ONB Labs) and cannot offer a broader service yet. Then again, while access to all interfaces is free in Finland and in Austria, in France users have to pay for access and use of Retronews, the advanced analysis tool of BNF.

Therefore, although the starting position is different for all interfaces and the outcomes vary accordingly (and it has to be stressed, that not all of the items mentioned below are missing on all interfaces, on the contrary, especially the websites designed for specialised analysis, Korp and Retronews, already have some of the desired features implemented) still some overreaching conclusion can be made.

First, in some cases the OCR quality is still very low. After identifying some major issues in this regard, the DH team's reliance on (and trust in) some search results was very low. This also seems a very important issue if it has to be assumed that users overall do not reflect on the quality of their search-hits and read the results as prove of their research.

Second, and intertwined with the OCR quality, both the surveys and the NewsEye-testing showed that keywords and context are key for success: topic and keyword suggestions could help users to achieve better results. Improvement of the search options would help especially academic users who readily use advanced search options, such as Boolean operators, asterisks, and other wildcards, if they are available. Advanced research cases would also welcome regular expressions in searches.

Third, for some research cases (such as genealogy, the history of places, cities or organisations, associations and clubs, or historical network analysis, etc.) tools that use named entity recognition (and possibly linking), which are especially important for family historians and chroniclers as well as for many academic analysis and visualizations. Optical

layout recognition, like article separation, image or advertisement recognition, on the other hand would allow for a targeted and complex analysis of easier to select corpora.
Fourth, the possibility of a personal workspace with functions to create subcorpora and annotation tools also gained the respondents' interest, confirming another assumption of the NewsEye DH team that underlined the academic need for such a working environment. The lack of designated personal workspaces on the main interfaces was especially felt as an obstacle for detailed qualitative analysis. Personal workspaces should include possibilities both to include and to exclude search results, to limit corpora, to annotate and to use advanced functions for analysis and visualisations. Closely linked to this personal workspace download options for newspapers or articles were asked for. Downloads of issues of newspapers in different formats (TXT, PDF, and JPEG) and of entire corpora (including metadata) are needed for further analysis. Also, download options e.g. of hits or of keywords in context (KWIC) would improve accessibility of the corpora.
Fifth, tools for analysis and visualisation both on the interfaces and in a personal working area greatly improve the understanding of performed searches and to contextualize results. Some interfaces, however, so far offer no tools to analyse and visualise search results and/or metadata. Among the tools requested and adressed were frequency analysis, KWIC analysis (keyword in context), co-occurrences, concordances, context visualisation, word clouds, graphs, publication periods, amount of issues, and, last but not least, geographical visualisations,. Simply put, text mining options would be welcome by newspaper researchers. This is also true for metadata. Metadata is lacking for many newspaper features, like metadata on the whole corpus (number of words, pages, languages etc. per year/month/day). There is also some need to correct (or to offer possibilities to indicate) errors in metadata. By improving download possibilities, some of these issues could be resolved, but many users outside academia would profit more from visualisations and possibilities for analysis on the interfaces themselves.
Sixth, frequency analysis and graphs are usually the next research approach for historians when trying to make sense of a large number of hits. Frequencies can help relate search results to historical processes by contextualising numbers that otherwise could lead to simplifications and erroneous assumptions. However, even though at first glance it seems trivial to create frequencies based on search results, there are things that need to be considered when working with them. Absolute frequencies can be skewed because they do not consider the whole dataset. On the other hand, in order to create relative frequencies, good metadata is needed. Producing frequency graphs is (computationally) not a demanding task, but downloading and processing digital data is still left in the hands of individual users or projects, a procedure prone to errors and misconceptions that can be very time consuming. Frequency analysis and graphs should therefore be integrated tools on the newspaper interfaces. Although libraries might struggle to offer appropriate functions for frequency analysis and graphs on their newspaper interfaces, it has to be taken into consideration that faulty frequencies can point to errors in the dataset, such as bad OCR quality, missing data, incorrect metadata, etc. To implement such a tool could therefore ultimately be to the advantage of the libraries.
Seventh, frequencies are also needed if topic modelling approaches are considered as alternative ways to analyse the datasets. The DH teams of NewsEye tested various topic modeling tools, some of them implemented on the newspaper interfaces, but most of them available on other sites. The tools tested with subcorpora from the case studies were STM, TidySupervise and Overview. The overall conclusions that can be drawn from the extensive testing of topic modelling tools are that many of the available tools are not flexible enough for what an increasing number of humanities researchers and a general audience need. To defer these challenges to (often wrongly assumed) user's inability to grasp complex issues is far too

simple. Users of newspaper interfaces need to be able to receive good results without computational skills, detailed knowledge about data structure and management or the basic functionality of topic models. As it is, results can often not be interpreted by users, because the computed topics do not make sense to the human eye. Most of the time, links to original data or documents are missing. Interpreting clusters of words can therefore be tiresome and frustrating. Historians need to see the context of chosen words in order to be able to understand why the tool selected certain terms and also in order to be able to be certain of the accuracy of the results. In addition, users have few options to interfere, even though different meanings of terms in the dataset, different contexts, word synonyms and bad OCR can influence the outcome of topic modelling tools. Topic modelling tools aimed at a general audience therefore should put user friendliness high on their agenda.

Eigth, it is not enough to create powerful tools: Their accessibility should be guaranteed by the mediation of professionals (for example librarians). At the same time, the dissemination and communication of the interfaces and tools developed was also a very important point in the surveys. The low representation of the category of students and those under twenty shows that information campaigns, educational material and example-pages about the value and use of digitized historical press among secondary school teachers and pupils, but also at universities is needed.

At the same time, and this was a rather surprising eight outcome of the surveys, many different user groups are mostly rather proficient at using the interfaces. Howevere, there is ample opportunity for improvements and this is where NewsEye could make a real difference, because the tools and functions that are being developed within the project can fill many gaps with regard to the enhancement of search functionalities and the quality of results. Testing the interfaces to the core and using some sample data gathered from the interfaces with publicly available toolsets the following can be read as a summary of academic users' wish lists that adds up to the outcome of the surveys.

As an overall assessment it can be underlined, that even for small research requests, the current process of selecting and extracting single articles, news items and search results is extremely time-consuming. Every step that the creators of digital newspaper interfaces could make, to reduce user's time for selecting, downloading, analysing and visualizing corpora and subcorpora would be greatly appreciated not only by the academic community. Ultimately, however, the feedback gathered for this paper also underlines [Yangs et al., 2011] conclusions:

> 'We have found that we can automatically generate topics that are generally good, however we found that once we generated a set of topics, we cannot decide if it is mundane or interesting without an expert.[. . .] We have come to the conclusion that it is essential that an expert in the field contextualise these topics and evaluate them for relevancy.'

To summarise, it can be concluded that the functions tested by the DH teams were especially useful to both generate hypotheses and to support hypotheses. It can also be said that digital tools combining quantitative macroanalysis (big data analysis) and qualitative microanalysis (exact reading) have some significant advantages over other methods. Especially when dealing with massive amounts of data, the sheer quantity makes the traditional practice of exact reading untenable and an inadequate method of evidence gathering. Analysing this data with traditional micro-analysis approaches often turns out to be impossible and researchers have the impression that they are missing out on important things. At the same time, the arguments vouching for quantitative approaches can be used against macro-analysis: By looking at the data from afar, crucial things could be missed. Agreeing with [Jockers, 2013], it must be said that the two ways of analysis, therefore, should and need to co-exist. Or, to once again stress the findings here, quantitative and

qualitative methods have to be closely intertwined. In that sense, it could also be said that the intertwining of computer science and humanities research can lead to more adequate digital tools and methods in the field. As such, NewsEye as a highly interdisciplinary project can be proof of the validity of this argument.

REFERENCES

Blevins C. Digital History's Perpetual Future Tense. *Debates in the Digital Humanities*. 2016. http://dhdebates.gc.cuny.edu/debates/text/77.

Brett M. R. Topic Modeling: A Basic Introduction. *Journal of Digital Humanities*. 2013. http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/.

Ehrmann M., Bunout E., and Düring M. Historical Newspaper User Interfaces: A Review. *WLIC proceedings*. IFLA (Athens), 2019.

Garfinkel S. Tweet, @footnotesrising, November 3, 2012. https://twitter.com/footnotesrising/status/264823621799780353.

Graham S., Weingart S., and Milligan I., Getting Started with Topic Modeling and MALLET. *Programming Historian*, 2012. https://programminghistorian.org/en/lessons/topic-modeling-and-mallet.

Herder, J. G. Auch eine philosophie der Geschichte zur Bildung der Menschheit. Beytrag zu vielen Beyträgen des Jahrhunderts. Hartknoch (Riga), 1774.

James P. and Steger M. B. A Genealogy of 'Globalization': The Career of a Concept. *Globalizations*. 2014;11(4):417-434.

Jarlbrink J., and Pelle S. Cultural Heritage as Digital Noise: Nineteenth Century Newspapers in the Digital Archive. *Journal of Documentation*. 2017;73(6):1228-43.

Jockers M. L. Macroanalysis: Digital Methods and Literary History, Topics in the digital humanities. University of Illinois Press (Urbana), 2013.

KB Lab. Newspaper Ngram Viewer. http://lab.kb.nl/tool/newspaper-ngram-viewerlive-demo.

Kemiläinen A. Nationalism: Problems Concerning the Word, the Concept and Classification, *Studia historica Jyväskyläensia*. 1964(3).

Kurunmäki J. and Marjanen J. A Rhetorical View of Isms: An Introduction. *Journal of Political Ideologies*. 2018;23(3):241-255.

Langlais P. C. Distant reading the French news with the Numapresse project: toward a contextual approach of text mining. *Numapresse*, 2019. http://www.numapresse.org/2019/02/07/distant-reading-the-french-news-with-the-numapresse-project-toward- a-contextual-approach-of-text-mining/.

Nelson R. K. Mining the Dispatch. 2019. http://dsl.richmond.edu/dispatch/pages/home.

Overview – Visualize Your Documents. https://www.overviewdocs.com/.

Papadimitriou C. H. et al. Latent Semantic Indexing: A Probabilistic Analysis. *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems – PODS '98*. ACM Press (Seattle/Washington), 1998. 159-168.

Pechenick E. A., Danforth C. M., and Sheridan Dodds P. Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution. *PLOS ONE*. 2015(10).

Silge, J. The Game Is Afoot! Topic Modeling of Sherlock Holmes Stories. 2018. https://juliasilge.com/blog/sherlock-holmes-stm/.

Underwood T. The Life Cycles of Genres. *Journal of Cultural Analytics*. 2016. http://culturalanalytics.org/2016/05/the-life-cycles-ofgenres/.

Yang T., Torget A., and Mihalcea R. Topic Modeling on Historical Newspapers. *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities.* Association for Computational Linguistics (Portland), 2011. 96-104. https://www.aclweb.org/anthology/W11-1513.