TS •> Tijdschrift voor tijdschriftstudies ISSN 1386-5870
DOI: https://www.doi.org/10.18352/ts.345 TS •> #38, December 2015, p. 61-78.

Publisher: www.uopenjournals.org. Website: www.tijdschriftstudies.nl

Filtering the "News": Uncovering Morphine's Multiple Meanings on Delpher's Dutch Newspapers and the Need to Distinguish More Article Types

LISANNE WALMA

l.w.b.walma@uu.nl

ABSTRACT

The current categorization of digitized newspaper archives prevents optimal use of the range of newspaper article types. Drawing examples from close reading research into the reputation of morphine in Dutch newspapers from 1880-1914, this article discusses how further specifying the article types can benefit different ways of historical research into digitized newspapers. Based on this discussion, the article argues that the interface of Delpher should allow researchers to search specifically for headings. Furthermore, the semi-automatic categorization of article types based on headings and matching newspaper sections should also be investigated, supported by crowd sourcing and niche sourcing.

KEYWORDS

digitized newspapers, Delpher, morphine, historical research, article categorization

INTRODUCTION

The digitization of historical newspapers allows newspapers to be comprehensively researched by historians.[1] In the past decade, digital archival projects such as Delpher, Chronicling America, Europeana, Trove, and the British Newspaper Archive have digitized a vast number of newspapers. Bob Nicholson deemed this mass digitization a 'digital turn', with both practical and methodological implications.[2] On the practical side, the online availability of newspapers has saved historians the time of travelling to the archives, and made it easier to compare multiple newspapers and show how newspaper

[1] Adrian Bingham, 'The Digitization of Newspaper Archives: Opportunities and Challenges for Historians.' *Twentieth Century British History* 21:2, June 1, 2010, 225-231 (226).
[2] Bob Nicholson, 'The Digital Turn.' *Media History* 19:1, 2013, 59-73 (66).

discourse has changed over time.[3] The option to perform keyword searches has fundamentally changed historical research because newspapers can now be analyzed bottom-up instead of top-down. Before the digital turn, historians would select a title and focus on newspaper sections and headings to locate the topic of interest. Full-text search now allows historians to locate discussions beyond newspaper front page and headings, such as in a letter or feuilleton. Cultural historians especially benefit from the digitization of newspapers because with keyword searches they can access a spectrum of articles that contain information about how a subject becomes embedded in daily culture.[4]

Despite the benefits of full-text searching, scholars have argued that the way digital archives present articles as single "hits" is problematic. A list of hits removes the articles from the internal context of the newspaper, such as the location of the article on the page and its structure.[5] Careful manual examination of each hit is often necessary to determine the relevance of the article, which is challenging when the number of hits is over a thousand.[6] Digital archives reproduce part of the newspaper's internal context by categorizing various article types. For example, most digitized newspaper archives, including Delpher, separate advertisements from other articles, thus opening up a new avenue for historical research.[7] Delpher also offers the possibility to filter digitized articles for images with captions and family notices. Most newspaper archives, including Delpher, group the remaining articles in one category, generally called "articles" or "news". I argue that there is a need for additional filters for digital newspaper archives. For example, Allen and Sieczkiewicz show that historical research can benefit from editorials extracted from digitized newspapers.[8] Unfortunately, editorials and other article types remain hidden in most archives in the overarching category of news/articles.

In this article, I show that the possibility to filter additional article types from digitized newspapers will contribute to the quality of research based on this source. I provide examples from my own research on the public reputation of morphine between 1880 and 1914 in the Netherlands. Morphine is an opiate painkiller. Opiates are relevant research subjects because they border medical and non-medical use: on the one hand, they are powerful painkillers; on the other hand, they also feature in stories about addiction and illegal trade. The public's perception of their use changes over time and this can impact doctors and pharmacists' handling of the substances. Between 1880 and 1914, ideas on what constituted medical and non-medical use of morphine were in

[3] Marcel Broersma, 'Nooit meer bladeren? Digitale krantenarchieven als bron.' *Tijdschrift voor Mediageschiedenis* 14:2, 2012, 29-55 (29).
[4] Nicholson 2013 (63).
[5] See for example, Gerben Zaagsma, 'On Digital History.' *BMGN - Low Countries Historical Review* 128:4, December 2013, 3-29 (26).
[6] Hinke Piersma and Kees Ribbens, 'Digital Historical Research: Context, Concepts and the Need for Reflection.' *BMGN - Low Countries Historical Review* 128:4, December 2013, 78-102 (101).
[7] For example in: Aleksandra Trtovac and Natasa Dakic, 'Bringing Historical Newspaper Advertisement into Research Focus - Europeana Newspapers Project.' *Преглед НЦД* 25, 2014, 2-10.
[8] Robert B. Allen and Robert Sieczkiewicz, 'How Historians Use Historical Newspapers.' In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem - Volume 47*, Silver Springs, MD: American Society for Information Science 2010, 1-4.

constant flux. Mid-nineteenth century morphine took the medical world by storm when the invention of the hypodermic needle greatly increased its painkilling potency.[9] However, with the increased potency and distribution of morphine by doctors and pharmacists came stories of overdose, poisoning, and addiction.

A full-text search of digitized newspapers can offer insight into the variety of ideas about the use of morphine circulating in public discourse. Schudson considers the newspaper to be a cultural actor: 'producers – and messengers – of meanings, symbols, messages.'[10] Studying newspapers can offer insight into cultural patterns circulating in the public at the time.[11] This is especially relevant for the early twentieth century, when newspapers played a vital role in the distribution of information to the public.[12] Studying the different representations of morphine in digitized newspapers can help answer the question how the boundaries of medical and non-medical use are formed. Therefore, I have reviewed national and regional newspaper articles between 1880 and 1914 on Delpher containing keywords for morphine. By combining this analysis with full-text search of medical trade journals, I discovered that pharmacists and doctors during this period sought ways to engage with the diverse meanings of morphine to win public trust.

My research greatly benefitted from the wide range of article types on Delpher. News stories revealed how various ideas about morphine circulated. Feuilletons and cultural reviews helped me to form a picture of how literary works referred to morphine. Finally, in letters, columns, and editorials, different medical actors came to the fore giving meaning to morphine by addressing the boundaries of medical and non-medical use. The number of hits resulting from this query (1,688) made it possible to manually read and classify the various article types.[13] However, queries that generate a higher number of hits require researchers to select articles for close reading on the basis of filters and samples. Some researchers also approach large numbers of articles with digital tools that locate topics or generate word frequency lists and present the results in a model for the researcher to further explore.[14]

In this article, I will show that researchers would greatly benefit from a more refined categorization of article types when approaching queries that generate a large

---

[9] For more information on changing discourses on morphine, see for example: Marcel de Kort, *Tussen Patient En Delinquent: Geschiedenis van Het Nederlandse Drugsbeleid*. Hilversum: Verloren 1995 (23); Timothy A. Hickman, '"Mania Americana": Narcotic Addiction and Modernity in the United States, 1870–1920.' *The Journal of American History* 90:4, March 2004, 1269–94; Virginia Berridge, *Opium and the People: Opiate Use and Drug Control Policy in Nineteenth and Early Twentieth Century England*. London: Free Association Books 1999.

[10] Michael Schudson, *The Power of News.* Cambridge, MA: Harvard University Press 1995 (18).

[11] Frank van Vree, *De Wereld Als Theater. Journalistiek Als Culturele Praktijk.* Amsterdam: Vossiuspers 2006 (15).

[12] Frank van Vree, *De politiek van de openbaarheid: journalistiek en publieke sfeer.* Groningen: Historische Uitgeverij 2000.

[13] This practice is often referred to as "close reading".

[14] This is called "distant reading", a variety of techniques such as topic modeling, named-entity recognition, and word clouds are available to help researchers approach large corpora of texts without reading them all. For more information see for example: Franco Moretti, *Distant Reading.* London: Verso 2013.

number of results. Firstly, I discuss the article types I encountered during my research. Secondly, I use these articles to demonstrate how the separation of article types is beneficial to historical newspaper research. Finally, I discuss several ways to separate the various article types and give suggestions for future research.

News

A morphine query from 1880-1914 using Delpher located mainly news articles (1,020 of the 1,688 in total).[15] Articles were marked "news" if they included short statements about current events at the time. Figure 1 shows additional information about the content of the news on morphine. I determined the categories in figure 1 in an iterative reading of the newspapers and secondary literature. The events generally addressed morphine subjects of addiction, poison, overdose, trafficking, or the medical qualities of the drug, for example as a painkiller or sleeping aid.

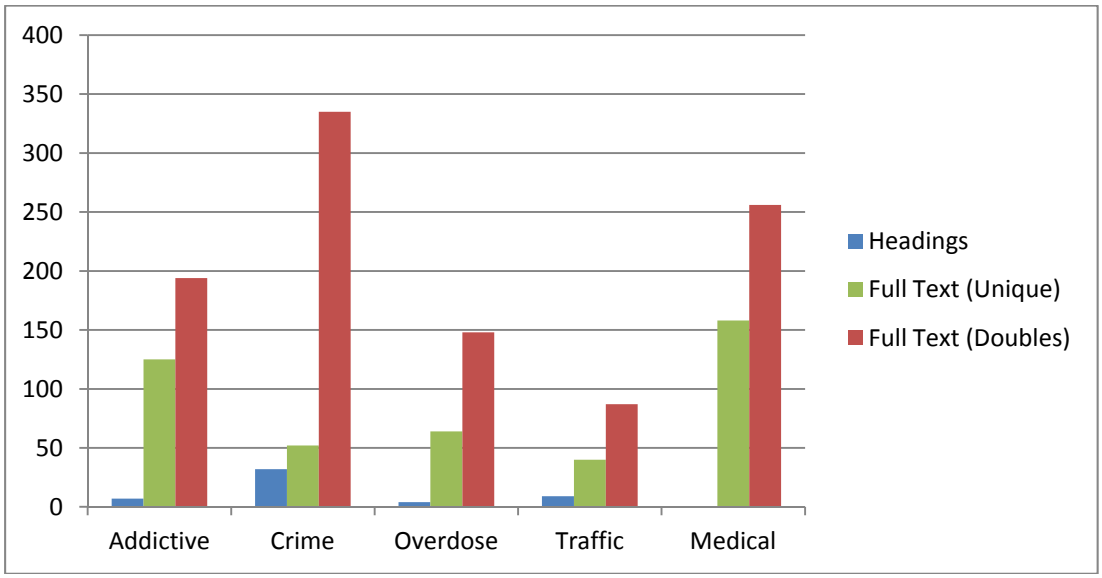

Figure 1: The various discussions of morphine based on searches of article titles versus full text (unique stories or including doubles) on Delpher.

A Delpher full-text search shows significant advantages over manual newspaper reading because of the increased number of results and article variability.[16] Figure 1 shows the morphine query outputs using headings versus a full-text search (with or without doubles).[17] The figure illustrates that a traditional research approach of skimming newspaper headings for the subject of interest would have identified only 52 articles. A full-text search, on the other hand, uncovered 1,020 news articles that mentioned morphine either in the text or heading. Moreover, figure 1 demonstrates the serial format of the news: a large number of articles repeated or followed up on the same story.

[15] This is based on a query using "morphine" OR "morfine" with filters: articles, national and regional newspapers.
[16] Nicholson 2013.
[17] Searching headings only is currently not possible on Delpher. I had access to a file containing the OCR of the articles and metadata, including headings, and generated these results with an Excel search.

Out of the 1,020 texts, only 439 referred to a unique news event.[18] Newspapers reprinted stories about crime in particular; these were mostly court cases.

The real problem with a search of just newspaper headings is that it provides a less diverse picture of morphine. Figure 1 illustrates that an analysis based solely on newspaper article headings would not have encountered the medical identity of morphine. In the results of a full-text search, I found stories on morphine as an important medicine, for example, as a painkiller for ill regents. Because these were articles that typically mentioned the use of morphine in passing, such as 'the doctors gave the regent morphine', these articles were not identified in the search of article titles.[19] We have seen how full-text search of the article type "news" showed a variety of settings in which morphine was discussed. However, research on morphine also benefitted from article types other than the news category.

BEYOND THE NEWS

In the search for newspaper articles for this research project, Delpher marked 1,688 texts as articles. Figure 2 shows that, although the majority of the articles were news-related, 450 articles could be categorized differently. The article types identified in figure 2 are a product of the research process and are not universal. Scholars could make other decisions. For example, some articles resembled news stories. Take background stories for example: due to their informative properties one could choose to categorize them as news. Other article types included feuilletons and letters, which deviated from the major news category and served other purposes; thus the decision to categorize them as separate seems less problematic. Moreover, article types are not a stable entity and their presence and content vary per period. Whereas today the International Press Telecommunications Council has defined specific genre classifications for news media, research into older newspapers demonstrates that their categories do not always fit these "modern" standards.[20] These issues will be addressed in the final part of this article. As for now, the article will focus on the specific categories identified during the period 1880-1914.

[18] Broersma 2012 (30).

[19] See for example: *Het nieuws van den dag: kleine courant*, 29 January 1889 (9).

[20] Robert B. Allen and Catherine Hall, 'Automated Processing of Digitized Historical Newspapers Beyond the Article Level: Sections and Regular Features.' In *Proceedings of the Role of Digital Libraries in a Time of Global Change, and 12th International Conference on Asia-Pacific Digital Libraries*. Berlin, Heidelberg: Springer-Verlag 2010, 91-101 (93).

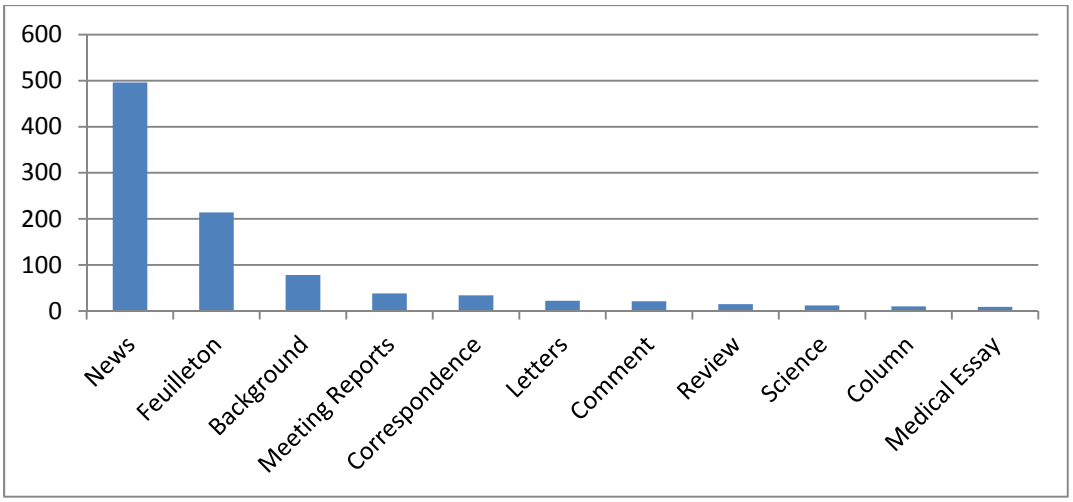

Figure 2: Various article types mentioning morphine, excluding doubles.

Feuilletons made up a large part of the results, whereas other types, such as letters, only delivered a handful of results. The category of news stories, the largest one, already showed a highly diverse picture of morphine, but the information drawn from the other article types contributed further to the analysis of newspaper articles mentioning morphine. For example, non-news articles provided a glimpse into the discourse on morphine beyond the intellectual elite, or they illustrated how morphine appeared in works of popular culture.

Below, I will use the results of the case study on morphine to illustrate that there are a number of ways in which analysis using digitized newspapers can benefit from the option to filter these articles from the general pool. Firstly, additional article type filters can help to find more contextual information about a subject. Secondly, they create opportunities for separate analyses and comparisons of the various newspaper genres. Thirdly, they can assist the researcher in exploring the corpus with topic models. Finally, they can lead to more refined frequency-list analyses.

CONTEXTUAL INPUT FOR CLOSE READING

Among the Delpher results were several letters to the editor in which pharmacists and drugstore owners fought over the right to distribute morphine and other substances. In 1898, for example, in a letter in *De Telegraaf*, a drugstore owner accused pharmacists of making mistakes in the distribution of morphine to patients.[21] There were also a number of angry letters by prominent pharmacists who accused drugstore owners of similar recklessness. For example, when a morphine poisoning case in the Netherlands was traced back to a drugstore owner who sold the perpetrator a kilogram of morphine, prominent pharmacist Leopold van Itallie sketched a disturbing image in order to convince the public of the dangers of letting drugstore owners handle the distribution of medical substances:

[21] *De Telegraaf*, 2 February 1898 (1).

> An important wholesale venue of medicine supplies a storeowner, without any medical authority, with one kilogram of morphine, an amount larger than most experts have ever seen in their lives. The storeowner subsequently makes this amount of morphine, which is enough to kill 2,500 people, available to a stranger. He does not even think to examine why this surprisingly large amount is being bought (…) He takes the same precautionary matters into account when distributing this poison as when he is selling a kilogram of sugar or chalk.[22]

Van Itallie called the drugstore owner reckless, questioned his expertise, and emphasized that his actions could harm a substantial number of people. These public fights between pharmacists and drugstore owners took place in letters, columns, and medical essays, but not in news stories. Combining these letters with a full-text search into medical trade journals showed that pharmacists were occupied with presenting their relationship with morphine in a positive way and publically blacklisted their competitors' handling of the drug. This finding added new layers to the story of morphine in the news; it showed different actors discussing the substance and using it as an example to promote their own professions.

Information about different actors discussing a topic in letters, essays, and columns can offer additional insight into public discourse and stakeholders. In this study, I was able to manually separate letters from the other articles. However, an analysis of a larger corpus may have had to rely on samples or distant reading. This would have prevented items such as the letters quoted earlier from coming to the forefront, as they represented only 23 of the total number of 1688 articles. The option to filter for letters, columns, background stories and editorials would be helpful, as researchers could choose to initially search for contextual information in a smaller subset. This would allow for a deeper understanding of the topic, rather than being constantly confronted with recurrent news headlines. If Delpher offered the option to filter for "letters" or "columns", I could have started my research there and located more information about the stakeholders involved in the debate on morphine.

Researching specific genres and questions

Since it is not yet possible to filter articles into smaller subsets, it is hard to acquire relevant newspaper texts that answer questions about the development of a specific genre or about how a specific subject circulated in spheres beyond the news, for example in works of popular culture. Feuilletons illustrate this point. They generated a substantial number of hits for morphine. During the period of this research, feuilletons mainly included serial short stories that were often translated from foreign languages into Dutch, since this was cheaper than writing an original piece.[23] Scholars have shown that the feuilleton provides potentially rich source material for analyses of the role of

[22] *Nieuwe Rotterdamsche Courant*, 14 November 1913 (3).

[23] Joan Hemels, 'Het feuilleton als fenomeen én fantoom in de Nederlandse journalistiek.' *Tijdschrift voor Tijdschriftstudies* 10:19, June 2006, 15-27 (16).

newspapers.[24] Researchers interested in the way the genre of the feuilleton emerged and developed could also benefit from a filter to more easily access this category.

Moreover, the option to pick and mix from the different article types when looking into a particular research subject can benefit studies driven by a specific research question, as some article types may be more relevant to the research question than others. Cameron Blevins has argued that reading all newspaper content together in one distant reading analysis prevents the reader from biasing one kind of type over others. This seems to be a correct observation, since readers might have been more interested in serial short stories than in the headlines or vice versa.[25] However, I argue that separating article types is a viable research strategy, not because one mode of information is inherently "better" than another, but because different types can show different aspects of a topic. For example, if a researcher was to be interested in how literary works dealt with a specific topic during the period of 1880-1914, it would be more logical to examine cultural reviews and serial short stories included in feuilletons first, before looking at news stories. This is exemplified by the way that feuilletons and literature reviews offer insight into literary works dealing with morphine.

Within serial short stories, the meanings of morphine were specific and recurrent. The stories focus on a reliable cast of characters that usually include either women addicted to morphine or at the end of their lives; their doctors prescribe the drug in both cases. Literature review articles from the studied time period also highlight this side of morphine. For example, critic Weruméus Buning complained that authors now felt it was stylish to poison their characters with morphine: 'There are authors that appear to think that the elevation of art means tying a rope around his [the character's] neck by the end of the story and hang him, or to give him a dose of morphine or chloral or another dirty substance.'[26] Buning's comment and a reading of the serial short stories in feuilletons suggest that works of literature, such as novels and short stories, dramatized the use of morphine.

If a researcher had the option to read the different article types separately, he or she could draw comparisons that reveal important aspects of their subject. For example, had I read all available articles that mentioned morphine I might have concluded on the basis of distant reading that crime was the most discussed subject since this was the most prominent news article type. If I had separated articles into a smaller subset including news, feuilletons, and reviews, I might have concluded that within literature, morphine mainly featured as an addictive and painkilling drug, whereas in news, its medical use featured prominently alongside crime.

[24] Korrie Korevaart, 'Teken van verandering: de opkomst van het feuilleton in de Nederlandse dagbladpers.' *Tijdschrift voor Tijdschriftstudies* 9:17, June 2005, 8–19; Hemels 2006.
[25] Cameron Blevins, 'Space, Nation, and the Triumph of Region. A View of the World from Houston.' *Journal of American History* 101:1, June 2014, 122-147 (128).
[26] A. Weruméus Buning, 'Zelfmoord-Epidemie Op Letterkundig Gebied.' *Het Nieuws van Den Dag*, February 19, 1893 (1).

In this study, the articles on morphine could be read manually. However, in queries that generate a substantial larger number of articles, computational techniques are often necessary alongside traditional methods to make selections or provide information to support an argument. Topic models locate topics in large numbers of texts based on how often words occur together.[27] Researchers who know the texture of the corpus they model generate the best output, because they are able to assess the meaning and relevance of a generated topic.[28] It would be beneficial for researchers using topic modeling to have the option to include or exclude specific article types to uncover new topics from a large body of texts.

Topic models often reproduce information that is already known. For example, the number and patterns of slavery ads in the Richmond *Daily Dispatch* from 1860-1865, which according to Robert K. Nelson was remarkably close to the actual ad pattern.[29] In my study on morphine, a topic model of articles on morphine similarly reproduced known information. As discussed earlier, the serial short stories that made up the category of the feuilleton featured a typical cast of characters – namely upper class women and their doctors. The word cloud in figure 3 demonstrates the recurrence of the characters.[30] There are multiple occurrences of Dutch words for female characters ("meisje", "vrouw", "mevrouw") next to the Dutch word for doctor ("dokter") in the feuilletons of the studied period. While the Dutch word for "man" ("man") was also mentioned, a close reading of these articles shows that they were mostly about women and their doctors.

[27] Shawn Graham, Scott Weingart, and Ian Milligan, 'Getting Started with Topic Modeling and MALLET.' *Programming Historian*, September 2012.
[28] Megan R. Brett, 'Topic Modeling: A Basic Introduction.' *Journal of Digital Humanities* 2:1, December 2012, 12-16.
[29] Robert K. Nelson, 'Mining the Dispatch.' http://dsl.richmond.edu/dispatch/pages/home.
[30] The word cloud was generated using *Voyant* (removed stop words) and can be found here.

Figure 3: A word cloud of the most frequent terms in short stories while searching for the word "morphine".

When feuilletons are not filtered from other article types, the recurrent keywords become part of a topic model of the entire corpus. Figure 4 shows topics with the word "dokter" ("doctor") found by a topic model of all 1,688 articles in the corpus versus one based on news articles only. When comparing the words in the left column of figure 4 to the word cloud in figure 3, it becomes clear that the topic found by topic modeling all article types together includes many of the words also present in the feuilletons. This demonstrates the strength of topic modeling: it identifies the particular article type as a separate topic within the topic model because of the specific words used. When a researcher enters these keywords into full-text search, most likely feuilletons will come to the forefront as output.

All 1,688 articles[31]	**News articles only**
zeide **dokter** kwam ging man vrouw zag oogen stond oogenblik kamer huis dien wilde hand	koning twee zeer des weder toestand bericht plaats **dokter** werden ontvangen vier ernstig zoodat einde

Figure 4: Occurrence of the word "dokter" in topic models of all articles on morphine and those articles indexed as news only.

In the right column, we see that selecting only news articles, thus excluding feuilletons, causes the word "dokter" to fall into a topic of "koning / toestand / einde" ("king / condition / ending"). This suggests that the topic of palliative care for royalty was part of reported news. In a topic model of articles that did not distinguish article types, this topic

[31] Model generated by Java Topic Modeling Tool. Settings were 15 topics, 200 iterations, and a proportion threshold of 0.05.

would have been lost. However, a close reading analysis of all the articles demonstrated that palliative care and royalty were highly relevant parts of the story of morphine during this period.

Nelson has remarked that 'topic modeling and other distant reading methods are most valuable not when they allow us to see patterns that we can easily explain but when they reveal patterns that we can't, patterns that surprise us and that prompt interesting and useful research questions.'[32] When exploring the corpus through topic models, users can change the settings of the topic models to see if new topics emerge. As demonstrated by figure 4, the decision to exclude article types also impacts topics found by the modeler. Therefore, this setting should be made available to researchers when they explore topics in digitized newspapers, since filtering article types renders different topic model outputs.

EXTRACTING ENTITIES AND MAPPING

Frequency lists such as word clouds and named-entity recognition are another form of distant reading analysis that could also be further refined by the option to filter additional article types. Using named-entity recognition (NER), a classifier recognizes separate entities in a text. Researchers can use this technique for instance to identify all locations or companies mentioned in texts. They can use locations identified by NER to make interactive maps to visualize change over time. Cameron Blevins has successfully used this technique to show how the *Houston Daily Post* produced space between 1894 and 1901.[33]

A part of my research is also concerned with the various geographic locations mentioned in newspapers. I particularly sought to determine how ideas from the United States featured in Dutch discourse on morphine.[34] Figure 5 shows the top ten locations found by NER in all articles. Dutch words which denote the United States were mentioned in news articles slightly less often than when all articles were analyzed together. This finding was based on a quick scan of the articles, and other factors also influenced the output of NER, such as the quality of the OCR or the applicability of the classifier for the specific language used in historical documents. However, the changes in relative numbers show that the inclusion or exclusion of article types changed the order of the geographic locations identified by the NER tool.

[32] Nelson.

[33] Blevins 2014.

[34] To learn more about this project, visit http://www.translantis.nl

All Articles[35]		News Articles		Feuilletons	
1. Amsterdam	341	1. Amsterdam	262	1. London	16
2. Rotterdam	241	2. Rotterdam	207	2. England	12
3. Paris	219	3. Paris	172	3. Paris	10
4. London	170	4. London	129	4. Bergen	8
5. Utrecht	135	5. Utrecht	102	5. Bern	7
6. England	130	6. Berlin	98	6. Europe	7
7. Berlin	127	7. England	97	**7. America**	**6**
8. America	**113**	8. Brussels	93	8. Schelde	6
9. Netherlands	111	9. Antwerp	89	9. Japan	5
10. Brussels	104	**10. America**	**81**	10. Halifax	5

Figure 5: Top ten geographic locations mentioned in all articles, news, and feuilletons.

The option to filter for additional article types allows researchers to ask more specific questions of frequency lists, enabling them for example, to look for specific locations, like "Amsterdam", in specific categories, like "news articles" or "feuilletons". They can then draw comparisons that tell them more about the discourse than when all articles are read together. For instance, extracting the locations in news articles on morphine suggests that events involving morphine, such as trafficking, court cases, poisonings, and addiction during this period mostly featured the Netherlands and Europe. Feuilletons, on the other hand, had a more international outlook than news articles, which is to be expected considering that the many short stories that made up the article type were translated.

An overview of locations based on NER, such as in figure 5, should always invite the researcher to do further close reading of the results to establish if the location mentioned is actually connected to the subject. Additional article type filters would ensure that researchers actually know on what kind of documents the maps and frequency lists they retrieve are based. This is important because different article types handle locations differently. In feuilletons, where the narcotic is often only mentioned in passing in a large body of text, the connection of a location to morphine may be less significant than in short newspaper articles, in which the location which is mentioned often refers to the actual location where an event involving morphine took place. In the NER based on unseparated article types, America emerges as the eighth most frequently mentioned location. However, this result may be attributable as much to news as it is to sports scores, a short story, or even a character named "America". The option to filter article types would allow the researcher to further refine these kinds of analyses.

[35] The table was generated using the SPSS IBM Modeler software with standardized Dutch library.

As mentioned earlier, newspapers do not offer a single type of article, but a mix of articles with distinct qualities. The range of newspaper articles that could be searched using Delpher proved to be invaluable for research on the public reputation of morphine in the studied time period. The various article types showed different actors mentioning morphine, literary works addressing morphine, and the international locations that featured in news about the drug. Additional article categories could be filtered. For example, studying articles mentioning heroin in the early twentieth century yielded recurring stories on a racing horse named "Heroin" in sports scores. While this topic might be interesting for researchers investigating the life of the *horse* heroin, these articles do not further research into discourses on the *drug* heroin. The option to filter out specific article type obviously aids distant reading analyses and helps the researcher to make a better selection for close reading. Should we then create a separate article type for sport scores or horse races? How far should categorization go? Currently, the researcher does have the option of using a Boolean term such as "NOT" to filter out these articles. This begs the question: Which article types should be filtered by Delpher and which articles can be manually filtered by researchers? In this final section, I explore this question, but first I give more information about how libraries generally separate article types.

Many libraries digitize newspapers through DocWorks, working together with Content Conversion Specialists. Optical Layout Recognition (OLR) is an important way to separate various article types during this process. OLR recognizes text blocks and lines on scanned pages, and can differentiate between illustrations, advertisements, and other texts.[36] This automatic article separation is combined with basic manual correction. Foreign contractors, who are often non-native speakers of the language used in the articles they are separating, correct the layout analysis and merge articles spanning multiple pages. These contractors are also responsible for article categorization, but since they do not speak the language used in the articles, this categorization is mainly based on the layout of the articles. In the case of Delpher, the archivists of the National Library of the Netherlands (KB) decided to distinguish between advertisements, family notices, images with captions, and articles.[37]

Because of the costs of separating article types, the KB should make choices about which types of articles to separate on Delpher. It should be remembered that the range of article types and their content changes overtime; thus there is also a need to establish in which periods newspaper content remain more or less stable. For the period researched in this article the focus could be on separating feuilletons, letters, and news articles. The headings of each of these three categories offer opportunities for automatic categorization. Because of the large presence and contextual impact of these three types of articles, the possibility to filter these articles will greatly advance both distant and close

[36] Claus Gravenhorst, 'Optical Layout Recognition (OLR) From Unstructured to Structured Newspaper Data'. ENP Information Day, November 2014.
[37] 'Verwerking.' www.kb.nl.

reading. The results of topic models and NER will become less cluttered on a distant reading level. On a close reading level, the option to filter for these article types makes it easier for the researcher to look beyond the headlines for public discussions and assumptions about a particular subject.

Some libraries already separate additional article types. Besides the possibility to select advertisements and family notices, the Trove project of the National Library of Australia offers researchers the option to select lists, results, guides, and literature in digitized newspapers.[38] Program manager Rose Holley writes that categorization depended on people who did not have English as their primary language, so articles that required comprehension beyond the first few lines were hard to categorize. They have done these categorizations via contractors in India on the basis of layouts and a "limited vocabulary list." Holley argues that a vocabulary list is helpful to categorize article types across a newspaper with known contents, headings, and layouts, but that the vocabulary changes over time and per newspaper. Adapting the list is too costly to do within the scope of the Australian newspaper project.[39]

Text-mining or topic modeling for recurring words can help establish vocabulary lists to distinguish article types over time. Allen categorized article templates based on the presence of specific words; for example, in weather reports, typical phrases like "rain", "sun" and "weather" are likely to be adjacent.[40] Refinement of these text-mining models could help automate the separation of article types. Some article types are relatively easy to separate depending on the quality of the OCR. Headings can particularly allow archives to distinguish among article types. For example, feuilletons usually have the header "Feuilleton", and in this period, because they consisted solely of serial short stories, the heading was followed by the author's name and the language from which the story was translated. Figure 6 shows a word cloud of the words most frequently mentioned in the titles of feuilletons found in the newspaper articles on morphine in the studied period.[41]

Figure 6: Word cloud of the headings for the article type "feuilleton".

[38] 'Advanced search - digitised newspapers and more.' http://trove.nla.gov.au.
[39] Rose Holley, 'Newspaper Article Categories'. National Library of Australia, May 2008.
[40] Allen and Hall 2010.
[41] The word cloud was generated using *Voyant* (removed stop words) and can be found here.

Because of the specificity of the headings for this article type, researchers can filter out feuilletons using Boolean search terms. Using the term "NOT feuilleton" in a query can remove most feuilletons from an analysis. Therefore, Delpher should include a "headings only" search field in the interface, because a title that mentions "feuilleton" is most likely going to be a feuilleton. Researchers interested in feuilletons or other sections of the newspaper with distinct headings would be able to acquire the material without the library performing a re-categorization.

Previous attempts to automatically classify article types were not entirely successful because of limited OCR-quality.[42] The advantage of using headings to filter instead of full text is that headings are often improved and re-keyed after digitization. However, the article type of letters shows that filtering by headings has its limitations. Letters during this period often had headings including the Dutch words "Ingezonden Stukken" ("Submitted Pieces"). Delpher classifies multiple letters printed back to back as part of separate sections. So while the first letter following the heading can be located by searching for 'Ingezonden Stukken', subsequent letters that are part of that section are not so easy to find. Figure 7 illustrates this problem.

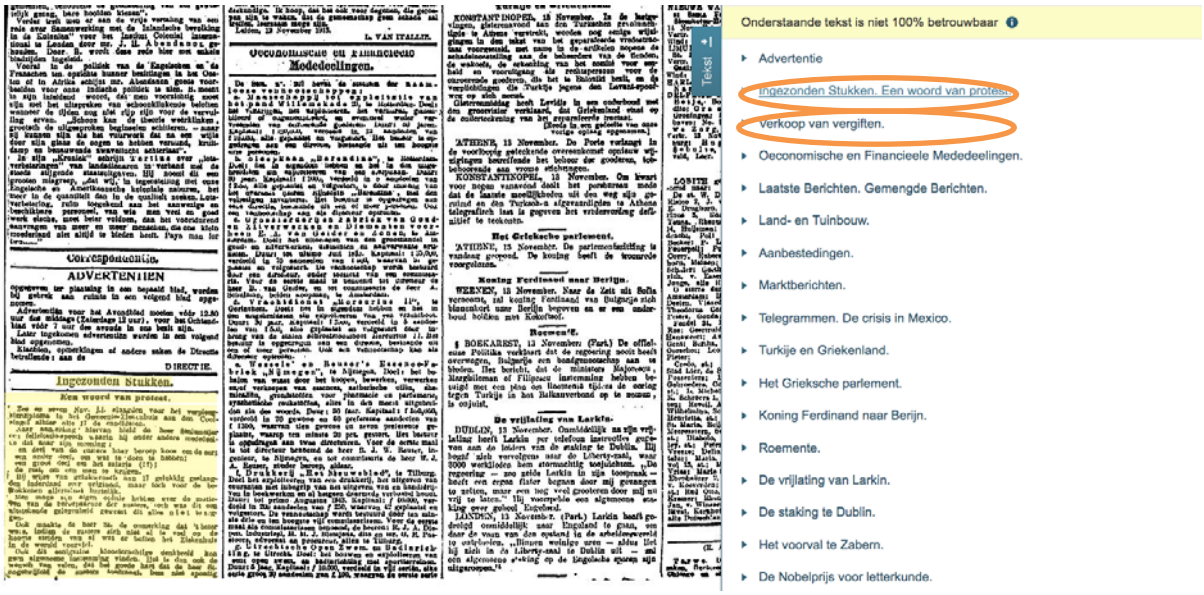

Figure 7: Two letters are circled that the interface of Delpher lists as separate sections even though they are both part of the letters section. The first letter features the recurring keywords "Ingezonden Stukken", while the other only features the title of the letter.

What would solve this issue is a more refined OLR which identifies separate articles, but also recognizes them as part of the same section. The large number of manual corrections that were necessary to mark articles spanning multiple pages will probably make the creation of a new index based on sections instead of articles a time-consuming endeavor.

[42] Marian Hellema, *Eindverslag Onderzoekstraject Tekstontsluiting*. Den Haag: Koninklijke Bibliotheek 2008 (8–9).

The issue is similar for the titles of news articles. The word cloud in figure 8 shows how these headings feature recurrent words such as "Gemengd Nieuws" ("Mixed News"), "Rechtszaken" ("Court Cases"), "Binnenland" ("National") or "Buitenland" ("International").[43]

Figure 8: Word cloud of the most frequent words in news article titles.

However, many news article titles also simply list the subject, thus articles under the header "news" are again marked as separate if they do not directly follow the header. One way to allow researchers to filter for news articles is to examine only the first page, which is already an option in the advance search of Chronicling America.[44] The validity of this approach to identify news articles depends on the period, which is being studied, since the nature of the front pages of newspapers has changed significantly over time.

Moving away from the specific period researched, another issue becomes apparent: the keywords that mark sections and the layout of the newspaper also change over time. As previously discussed, modern thesauri are not useable for older articles, such as articles from the early twentieth century.[45] Thus, it appears that categorization requires manual corrections and user input. Libraries may be hesitant to start this process because of the large amount of time and number of resources required. Edwin Klijn spoke about the long process of newspaper digitization: 'Segmenting a newspaper at article-level can be time-consuming; the processing speed is estimated by one respondent at approximately 100 pages per hour. At that rate, to segment all 8 million newspaper pages for the DDD project to page-level would take one employee about 50 years!'[46]

There is no denying that article separation requires a significant investment, but in the end everyone stands to benefit from improved filtering of the large number of

[43] The word cloud was generated using *Voyant* (removed stop words) and can be found here.
[44] 'Advanced search.' http://chroniclingamerica.loc.gov/.
[45] Hellema 2008 (8–9).
[46] Edwin Klijn, 'The Current State-of-Art in Newspaper Digitization: A Market Perspective.' *D-Lib Magazine* 14:1/2, January 2008.

digitized newspaper articles. Libraries and researchers should work together to allow this separation; there are opportunities for crowd sourcing and so-called niche sourcing.

Libraries have made good use of crowd sourcing. Trove, for example, allows users to correct OCR mistakes. Niche sourcing, which is when experts are asked to provide knowledge on a topic of their expertise, is a welcome addition to the improvement of digital newspaper archives.[47] The crowd can do part of the work for the article type separation, such as matching articles if they are part of the same section and correcting the heading texts. Since type-specific keywords, newspaper layouts, and the article types present in newspapers tend to change over time, researchers conducting newspaper analysis in libraries should share their remarks and categorizations with the digital archives as part of niche sourcing. Drawing on researchers' expertise and experience makes it easier to point out recurrent words that mark a section for a particular period. Researchers can describe the characteristics of sections and use them to automatically identify article types. The Delpher team therefore should remain in close contact with the various groups that are currently using their newspapers, since these researchers hold valuable information on the content and form of newspapers in their collection. A working interface in which researchers could tag particular sections, and an active communication between libraries and digital historians on social media would allow the library to engage with these users.

Conclusion

This article has shown that the categorization of most articles in digital archives today as one type of article prevents the optimal use of digitized newspapers for all researchers. This affects researchers exploring a corpus through close reading as well as those using distant reading. Filtering additional article types would help researchers investigate large numbers of digitized newspaper texts in several ways. Firstly, the option would enable researchers to locate articles that may be overlooked in distant reading, but that are highly relevant for answering their research question. For example, letters may offer a wider perspective on the different actors involved in discussions on a particular topic compared to newspaper headlines. These contextual articles could serve as a starting point for close reading. Secondly, filtering article types would allow for the analysis of the development of specific genres printed in newspapers, such as feuilletons. It would also allow the researcher to pick and mix source material that is relevant to the research question, such as feuilletons and cultural reviews if circulating literary works are considered. Thirdly, filtering article types would be another option for researchers to experiment with using topic models to explore newspaper discourse. Finally, article separation would refine the input for frequency-count analyses, such as NER and

[47] Jasper Oosterman et al., 'Crowd vs. Experts: Nichesourcing for Knowledge Intensive Tasks in Cultural Heritage.' In *Proceedings of the 23rd International Conference on World Wide Web*, Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee 2014, 567–68.

mapping, because the researcher could ensure that frequency lists were based on similar articles, or at least know the background of the material analyzed.

Vocabulary lists based on headings offer researchers the option to filter articles manually, and they allow libraries to consider the automatic classification of article types. The results of this research suggest that Delpher should allow researchers to search for headings only in the interface while considering a semi-automatic categorization of article types based on headings and sections. This approach would require a significant effort, since it would most likely involve some form of manual correction. However, an investment in article categorization today will greatly enhance research in the future. Moreover, Delpher and other digitized archives can benefit from cooperation with crowds and experts. Crowd sourcing helps to refine OLR recognition by connecting articles to overarching sections, and niche sourcing helps to establish how vocabulary lists, newspaper layouts, and article types have changed over time.

What article types researchers need to filter depends on their individual research questions and chosen tools, but there is no doubt that the option to pick and mix article types would add significantly to the exploration of the ever-growing corpora of digitized newspapers.

•> LISANNE WALMA *is a PhD Candidate at Utrecht University. She is part of the Translantis project, in which researchers use text-mining to study the role of the United States in twentieth-century Dutch public discourse. More information on this project can be found on* http://www.translantis.nl.