Can Umlauts Ruin Your Research in Digitized Newspaper Collections?

A NewsEye Case Study on 'The Dark Sides of War' (1914–1918)

Barbara Klaus

Department of Contemporary History, University of Innsbruck, Austria
barbara.klaus@uibk.ac.at

Abstract. Digitized newspaper collections facilitate the access to historical newspapers. Even though they offer several useful possibilities regarding the research in historical newspapers and magazines, the (automatic) research in these collections is (still) full of limitations and pitfalls. Based on the research conducted on the platform AustriaN Newspapers Online (ANNO) for the NewsEye case study 'the dark sides of war', the main challenges of working with digitized newspaper collections will be discussed in this paper. Especially two aspects – the fire catastrophe at the munitions factory Wöllersdorf (1918/09/18) in Lower Austria and the Austrian press coverage about war widows during the First World War – will be used as specific examples. The discussed limitations include the Optical Character Recognition (OCR) quality, provided search options and metadata, as well as others. Furthermore, possible improvements regarding these challenges, e.g. Optical Layout Recognition (OLR), Named-entity Recognition (NER) and Named-entity Linking (NEL), will be presented in this paper.

Keywords: Digital Humanities, First World War, Digitized Newspaper Collections, Historical Research Interfaces.

1 Introduction

In an increasingly digitized and globalized world, people expect that all information is available online. This idea also includes historical data, which was never originally intended to be available on the World Wide Web. However, the ongoing digitalization process in the field of historical newspapers has to be considered as one expression of the current attempt to make old issues of the mass medium available for the general public. Today, many people all over the world – researchers, librarians, students and others – use digitized newspaper collections for very different reasons every day. These include, for example, scientific research, homework or actual 'reading' in their leisure time.

However, not all of these people are aware of the many possibilities such collections offer. On the other hand, as already mentioned, these newspapers were never intended to be browsed using digital methods at the time of their publication. As a result, the (automatic) search is full of limitations and pitfalls that have to be – if possible –

avoided. It can be assumed that irregular and even some regular users of digitized newspaper collections are unaware of this circumstance. In this paper, findings and experiences made during the research process for the NewsEye case study 'the dark sides of war' (1914–1918) in ANNO (AustriaN Newspapers Online) will be used as a basis to underline and illustrate some of these challenges.

2 ANNO – What or Where Is My Data?

The open-source platform ANNO (AustriaN Newspapers Online) is an online portal for (mainly Austrian) digitized historical newspapers and magazines provided by the Austrian National Library (ONB). It offers free access to millions of newspaper pages published between 1568 and (currently) 1948. The collection is being continuously extended with the newspapers and magazines published 70 years ago. [1] All users who are working with ANNO and other digitized newspaper collections have to face two different questions first: What is my data? Where is my data?

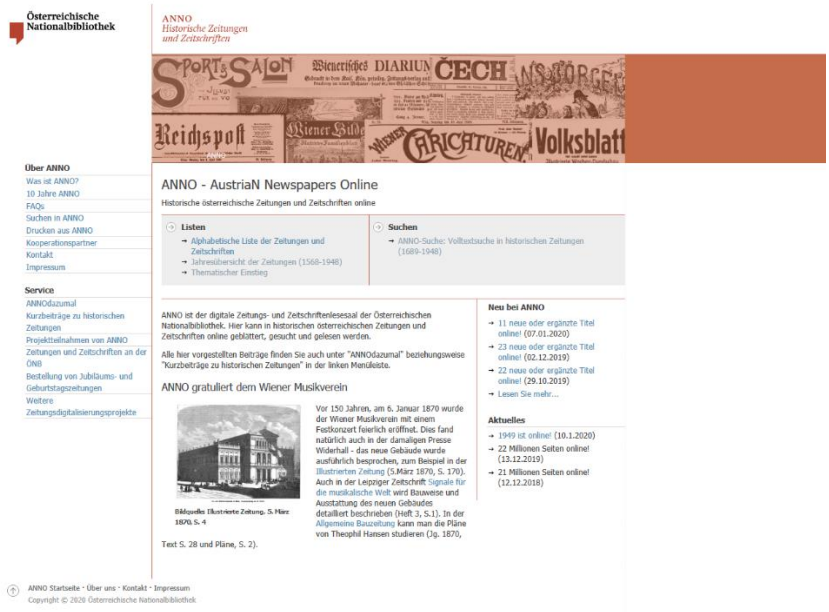

Fig. 1. Homepage of ANNO. It includes the entry to the digitized newspapers and magazines via lists ('Listen') or search options ('Suchen'). © 2020 Österreichische Nationalbibliothek.

Basically, it is possible to access the historical newspapers and magazines in ANNO by lists or through search options (see fig. 1). The first option, which includes an annual overview and an alphabetical list of the available newspapers and magazines, is especially useful as an entry portal for manual research or 'browsing'. The latter addresses users who already have a subject in mind that they are looking for. ANNO offers two different search options: a simple search and an advanced search. As the descriptions

already indicate, the advanced search allows a more refined search by language, year/date of publication, place of publication and newspaper or magazine. Furthermore, the so-called distance search enables users to search for two terms that are placed in vicinity of each other. The 'distance' – the number of words between both searched terms – can also be defined. In some cases, this is a very useful feature and makes a more targeted search possible.

ANNO provides single newspaper issues and pages in three different formats: PDF, JPEG and TXT. As this already suggests, there is no article separation implemented in ANNO yet. If you want to work with single articles, the text of an article has to be copied from the TXT document to another, for example, DOCX document – manually, of course. This very time-consuming process is currently the only way to get access to single articles in ANNO. Consequently, hours and days were spent copying long and short articles from ANNO to single TXT files or DOCX documents. Moreover, there is no personal working area, where a corpus could be compiled online. As a result, the management of corpora mainly takes place offline. Furthermore, no analysis and visualization tools are offered in ANNO. Therefore, other ways to evaluate and visualize the data have to be determined. Programs and online tools, like *Overview*[1], *Voyant Tools*[2] and *Wordle*[3] (see figure 2), offer viable alternatives. In the following paragraphs, further limitations regarding the research in ANNO will be discussed in detail on a more practical level.

3 The 'Dark Sides of War' (1914–1918) – A NewsEye Case Study

In the interdisciplinary and multilingual Horizon 2020 project "NewsEye. A Digital Investigator for Historical Newspapers" (2018–2021), DH researchers are working on different case studies. The present case study on the 'dark sides of war' (1914–1918) was developed and conducted during the project and therefore further subdivided into smaller, more practicable research issues. Especially two of them included considerable challenges in the research process: the press coverage on the fire catastrophe at Wöllersdorf and war widows during the First World War. The challenges were mainly caused by the research topics themselves and the limitations given by the online historical newspaper archive ANNO. These will be discussed later on. However, the already mentioned topics of research were also chosen for a completely different reason.

Between 1914 and 1918, the First World War developed into the first 'total war' and affected all parts of society. Consequently, the civil population was also involved in warfare more than ever before and this circumstance lead to some profound changes in society. [2] These included the emergence of a so-called 'home front' as a pendant to the war front in Austria and other warring countries. [3] Several negative impacts accompanied the war, which can be defined as the 'dark sides of war'. Destruction, death,

[1] https://www.overviewdocs.com/
[2] https://voyant-tools.org/
[3] http://www.wordle.net/

invalidity, war widows as well as war orphans can be added to a much longer list of terrible effects of war. At the same time, the First World War can be defined as the first 'media war', because all means of communication were used for propagandistic and patriotic purposes and suffered from censorship at the same time. [4] These circumstances have to be taken into account when working with newspapers published between 1914 and 1918. It was therefore decided to investigate two specific aspects – the accident in Wöllersdorf and the presence of war widows in the press – to underline the consequences of war for individuals far away from the battle front and to draw attention to these underrepresented and almost forgotten groups of victims.

3.1 Wöllersdorf – The Unknown Place?

Less than two months before the end of the First World War, the largest civil disaster in Austria during the conflict took place in Wöllersdorf, a town in Lower Austria. [5] It was also the most severe fire disaster in the 20th century in Austria. [6] The fire catastrophe at the local munitions factory claimed the lives of over 400 workers – only 35 people survived. [5] Manly young girls and women, who had to replace the male population in civil workplaces [7], died in this accident. [5] The fire in the building No 143 on the 18th of September 1918 was caused by a cartridge case that fell to the floor and rapidly set fire to the surrounding gunpowder. [6] The extremely high number of victims was caused by locked side exits, which were supposed prevent the laborers from leaving the building early for lunch. Due to this fact, the fire in one of the most important munition factories of the Austro-Hungarian arms industry was repeatedly compared to the fire in the Ringtheater (Vienna), which (officially) claimed 384 victims in 1881. [8]

First, the military administration tried to minimize the damage in public opinion and only allowed the printing of a short official statement. Later, an insight in the true extend of the tragedy was given, but – due to censorship – the official entities still had a huge influence on the press coverage about the fire catastrophe. [6] These interferences lead to so-called 'white spots', which made the impact of censorship visible for the general public. [9] A cumbersome manual research lead to the result that, over a time period of two weeks, 103 articles in 37 different Austrian newspapers reported on the accident in Wöllersdorf. Already after the first week, fewer and fewer reports were published. Even though it could be argued that officials prevented further articles about the fire catastrophe, it is more likely that, towards the end of World War One, other news were more relevant. [15]

Table 1. Keywords and hits (advanced, distance search).

Keywords	Hits
"Brand Wöllersdorf" ~ 20	10
"Brand Munitionsfabrik" ~ 20	41
"Brand Katastrophe" ~ 20	6
"Katastrophe Wöllersdorf" ~ 20	8
"Munitionsfabrik Wöllersdorf" ~ 20	12

However, it was a difficult task to prove the initial assumption that a lot of different newspapers reported about this serious accident. Several queries with different key words, such as 'fire' ('Feuer'), 'catastrophe' ('Katastrophe'), 'munitions factory' ('Munitionsfabrik') and 'Wöllersdorf', and combinations led to a small number (43 articles) of relevant results (see table 1). For this reason, a manual research in about 50 newspaper issues per day over a time period of two weeks (1918/09/19–1918/10/02) was conducted to validate the results. This manual research showed that more than 100 articles fitting the research question were published. Of course, the **OCR quality** is always an issue. The OCR quality of some articles was so bad that none of the keywords mentioned above were spelled correctly. Therefore, it was impossible to detect them with the full-text search.

Beside this, another factor distorted the results: Wöllersdorf itself. The place name was mentioned several times in the articles and headlines, but leads to few results, because it was misspelled most of the time. In this case, the sometimes lower OCR quality coincided with the circumstance that umlauts (and headlines) are especially prone to bad OCR recognition. Therefore, the accident happened at many places, like 'Wölkersdorf', 'Wollersorf' or 'Wallersdorf', but seldom Wöllersdorf. However, in ANNO, question marks cannot be used for umlauts. As a consequence, it is nearly impossible to effectively search for terms including umlauts in ANNO and there was no chance to get satisfactory results in the advanced search regarding the tragedy at the munitions factory Wöllersdorf. Also, the usage of asterisks or wildcards was not possible, because they do not work in combination with Boolean operators (AND, OR, NOT) in the advanced search.

In such a case a further development of the provided **search options** is necessary to improve the usefulness of the full-text search. The utilization of question marks for umlauts would mark a first step in the right direction. In this context, it has to be mentioned that umlauts are regularly used in German. Therefore, this problem concerns several research topics. Based on first-hand experiences, the umlaut issue especially involves Named Entities, like surnames (e. g. Müller, Hötzendorf and Dürer), companies and brand names (e.g. ÖBB, ÖAMTC and Jägermeister) and place names (e.g. Gmünd, Schönau and Kitzbühel). Consequently, the implementation of **Named-entity Recognition** (NER) would help to simplify the research. Furthermore, the circumstance that several places and persons carry the same name has to be taken into account. This problem can be resolved with the implementation of **Named-entity Linking** (NEL). Both, NER and NEL, require a good OCR quality and allow for a much more focused research process.

3.2 Widow ≠ Widow

First, the term 'war widow' has to be specified. It should be clear that war widows are those women who lost their husbands, serving in the army, to the armed conflict. [10] Their children, who lost their father during the war, are commonly described as war orphans. [11] Therefore, it has to be noted that there is a major difference between women who lost their husbands during and outside wartime. While traditional widows are expected to be old(er), war widows were usually young and in the prime of their

lives. In this context, Smith notes that the prefixing of the noun ‘war’ illustrates the superior status of war widows towards the traditional widows. This is highlighted by the fact that war widows, in contrast to traditional widows, received pensions by the state. Subsequently, as Smith states, war widows became ‘public property’ and, as a consequence, became more visible than traditional widows. [12]

Because of the so far unprecedented number of victims, the First World War led to a challenging number of invalids, war widows and war orphans. Post-war figures, cited by Hämmerle, indicate that approximately 350.000 war widows and orphans in Austria relied on state support during and after the conflict. [13] Winkelhofer divides these two groups of war victims and assumes that fallen Austrian soldiers left behind 90.000 to 95.000 widows and about 270.000 orphans. [14] These numbers can only partly outline the destructive dimension of World War One. [13] As a consequence, the resulting needs of the surviving dependents of fallen soldiers during and after the war led to an increased attention to the fate of war widows. [10]

Fig. 2. Most important words in the press coverage about war widows (*Arbeiter-Zeitung*, 1914/1918, created with *Wordle*).

According to this, it can be argued that war widows and their fates should have been a major issue in the newspapers between 1914 and 1918. Some queries in the advanced search lead to plenty of results. The fact that, in the press coverage, war widows were often only called ‘widows’ and not ‘war widows’ presents the central challenge in this research. Therefore, the search for ‘war widows’ supplied just a small part of all relevant results. Even the already mentioned distant search offered in ANNO was not useful in this case, because combinations like ‘widow’ (‘Witwe’) and ‘war’ (‘Krieg’), ‘front’ (‘Front’), ‘soldier’ (‘Soldat’) or ‘death’ (‘Tod’) led to few results. This fact can only be explained with the circumstance that widows are either rarely mentioned in this context or – again – **the OCR quality** was an issue. An in-depth analysis of the material has shown that the latter is the case.

However, some (reliable) results had to be delivered. Hence, a more broad approach was chosen for the query in ANNO. Even though most of the results with the single

term search ‘widow*’ were useful, it is not always easy to identify the ‘correct’ widow, because the system returns all kind of ‘widows’, not only war widows. As a result, the hits had to be filtered manually. In this context, a function to exclude unnecessary hits – or at least all hits in the ads – would be a huge time-saver. There are several ways to put this into practice. First, the possibility to include or exclude specific contents of newspapers (e.g. articles, pictures and ads) in the **search options** itself, would help to mitigate this problem. This requires the implementation of **OLR**.

The provision of a **personal working space** could be another way to face this problem. Even though the results still have to be checked manually, fitting articles could easily be transferred into this personal working space with a simple click. Consequently, corpora could be compiled online, which currently takes place offline. Another aspect in this context is that the hits are not sorted by date, but by the amount of hits in one newspaper issue. As a consequence, without a workspace to save the results, it is easily possible to lose track, which was (nearly) the case during the research on war widows. Furthermore, a personal working space could include some **analysis and visualization tools**. For example, word clouds are helpful to get a quick insight into the content of corpora on a quantitative level (see figure 2). The provision of a summary and further information (**Metadata**) regarding the compiled corpus in the personal working area would also simplify the creation of larger corpora and therefore significantly increase the usability of online platforms like ANNO.

4 Conclusion

With regard to digitized newspaper collections and especially ANNO, it should become clear that there is still significant potential for improvement in terms of the preparation of the material and the offer of editing functions. These provide an improvement of the OCR quality and the search options, the implementation of OLR, NER and NEL as well as the supply of a personal working space, which includes metadata as well as analysis and visualisation tools.

Due to the facts given, it also has to be noted that the reliance on automatically generated results has to be questioned and the results always interpreted with caution. Developments in the already mentioned areas are therefore critical to improve and facilitate the research in digitized newspaper collections for all user groups, such as students, lay historians and scientists. Fortunately, the currently planned relaunch of ANNO in 2020/21 promises some new developments, even though no further details can be published at the present.

Acknowledgements

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

References

1. ANNO – AustriaN Newspapers Online, http://anno.onb.ac.at/, last accessed 2020/01/01. Flemming, T.: Grüße aus dem Schützengraben. Feldpostkarten im Ersten Weltkrieg aus der Sammlung Ulf Heinrich. be.bra Verlag, Berlin–Brandenburg (2004).
2. Krumeich, G.: Kriegsfront – Heimatfront. In: Hirschfeld, G., Krumeich, G., Langewische, D., Ullmann, H.–P. (eds.): Kriegserfahrungen. Studien zur Sozial- und Mentalitätsgeschichte des Ersten Weltkriegs. 1st edn. pp. 12–20. Klartext-Verlag, Essen (1997).
3. Weiss, W.: "…den andern drauf Gusto zu machen. " Das Schlachtfeld der visuellen Kriegsdarstellung. In: Riegler, J. (eds.): "Ihr lebt in einer großen Zeit,…" Propaganda und Wirklichkeit im Ersten Weltkrieg. pp. 51–65. Steiermärkisches Landesarchiv, Graz (2014).
4. Linke, R.: Wöllersdorf 1918: 423 Tote klagen an. News ORF Homepage, https://noe.orf.at/v2/news/stories/2938436/, last accessed 2020/01/01 (2018).
5. Sabitzer, W.: Flammen in der Munitionsfabrik. In: Öffentliche Sicherheit 9–10/18, pp. 42–43, BMI Homepage, https://www.bmi.gv.at/magazinfiles/2018/09_10/brandkatastrophe.pdf, last accessed 2020/01/01 (2018).
6. Hammer–Luza, E.: "An den Schmerzenslagern unserer verwundeten Krieger." Die Krankenschwester im Ersten Weltkrieg – Ideal und Realität. In: Riegler, J. (eds.): "Ihr lebt in einer großen Zeit,…" Propaganda und Wirklichkeit im Ersten Weltkrieg. pp. 171–178. Steiermärkisches Landesarchiv, Graz (2014).
7. Simhofer, D.: Brandkatastrophe vom 18. Sept. 1918 in der k. u. k. Munitionsfabrik Wöllersdorf. Mein Bezirk Homepage, https://www.meinbezirk.at/wiener-neustadt/c-lokales/brandkatastrophe-vom-18-sept-1918-in-der-k-u-k-munitionsfabrik-woellersdorf_a2901119, last accessed 2020/01/01 (2018).
8. Bürgschwentner, J.: Propaganda. In: Kuprian, H., Überegger, O. (eds.): Katastrophenjahre. Der Erste Weltkrieg und Tirol. pp. 277–303. Universitätsverlag Wagner, Innsbruck (2014).
9. Kuhlman, E.: Of little comfort. War Widows, fallen soldiers, and the remaking of nation after the Great War. New York University Press, New York (2012).
10. Pawlowsky, V., Wendelin, H.: Die Wunden des Staates. Kriegsopfer und Sozialstaat in Österreich 1914–1938. Böhlau Verlag, Wien (et al.) (2015).
11. Smith, A.: Discourses Surrounding British War Widows of the First World War. Bloomsbury Publ., London (2012).
12. Hämmerle, C.: Heimat/Front. Geschlechtergeschichte/n des Ersten Weltkrieges in Österreich–Ungarn. Böhlau Verlag, Wien (et al.) (2014).
13. Winkelhofer, M.: So erlebten wir den Ersten Weltkrieg. Familienschicksale 1914–1918. Eine illustrierte Geschichte. 2nd edn. Amalthea Signum Verlag, Wien (2013).
14. Klaus, B.: The fire of Wöllersdorf (1918) in Austrian Newspapers. NewsEye Homepage, https://www.newseye.eu/blog/news/the-fire-catastrophe-in-woellersdorf-1918-in-austrian-newspapers/, last accessed 2020/01/05 (2019).