Testing Hypotheses with Dirty OCR and Web-Based Tools in Periodical Studies[1]

Malte Lorenzen

Periodical studies always played only a marginal role in the philologies. But for a couple of years there has been at least a small boom in research on journals and magazines which depends to a large extent on the emergence of the digital humanities.[2] The opportunity to browse through large corpora and search with ease for the rise and fall of distinct topics, or even for the use of single words over time made it once again attractive to deal with a type of text which was formerly often rejected as object for research because of its various contents, its ephemerality and its apparent distance from high literature.

The necessary condition of any research with the means of the digital humanities is the availability of a digital corpus that can be processed by computer tools. Yet this availability is still one of the most urgent problems. Though much work in the process of digitizing hundreds of years of print culture has been done, there is no guarantee that the very texts one needs for a project are available anywhere in the World Wide Web.[3] And even if

1 I am very grateful to Kai Kauffmann for our discussions on the possibilities and restrictions of tools and to Christine Peters and Joris C. Heyder for their critical readings of this article. Special thanks go to Laura Säumenicht for the digitization of the examined corpus.

2 A more or less initializing text for the new interest in periodicals as an object of the (digital) philologies is: *Latham, Sean, Scholes, Robert*, The Rise of Periodical Studies, in: *PMLA* 121 (2006), 517–531.

3 Cf. *Hahn, Carolin*, Forschung benötigt Infrastrukturen: Gegenwärtige Herausforderungen literaturwissenschaftlicher Netzwerkanalysen, in: Toni Bernhart et al. (ed.), Quantitative Ansätze in den Literatur- und Geisteswissenschaften: Systematische und historische Perspektiven, Berlin and Boston: de Gruyter, 2018, 315–34, esp. 326–28.

someone has already digitized the corpus in question, some of the texts might be nothing more than "a garbled mess" because of bad optical character recognition (OCR)[4] – the best known technique to bring printed texts into a machine readable form – and/or the condition of the underlying historical documents.[5] Not least because digitization is actually a very time consuming and, therefore, expensive process, some of the recent discussions in the digital humanities aim not only at ways to improve OCR but also at the question "how accurate […] digitized sources [must] be to produce robust results".[6] In other words, the discussion is about the possibility to obtain new scientific insights while working with 'dirty' OCR full of errors.

Previous studies were quite optimistic in this respect.[7] So as I am bound to a quite large corpus of documents in a research project on comparisons of cultures in German periodicals during World War I, I decided on digitizing at least a subcorpus to see how far I can get with it despite possible OCR errors. The main interest of this article is to map out the potential of provisional digital documents. Thus, this contribution has a somewhat experimental character because it not only rests upon dirty OCR but also upon capabilities of tools.

Apart from the availability of digital documents, the decision on software poses the most urgent problem for any digital philologist – even more so for any 'newbie' in Digital Humanities. When there is no research community including specialists in computer sciences who can develop software exactly fitting the interests in a given corpus of documents, it is the best option to choose free available tools and toolkits like *Voyant Tools*[8] or *AntConc*[9] with a

4 Cf. *Nicholson, Bob*, Counting Culture; or, How to Read Victorian Newspapers from a Distance, in: Journal of Victorian Culture 17 (2012), 242.

5 Cf. *Cordell, Ryan*, 'Q i-jtb the Raven': Taking Dirty OCR Seriously, in: Book History 20(2017), 194–95 and *Holley, Rose*, How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs, in: D-Lib Magazine 15 (2009), https://doi.org/10.1045/march2009-holley.

6 *Strange, Carolyn et al.*, Mining for the Meanings of a Murder: The Impact of OCR Quality on the Use of Digitized Historical Newspapers, in: Digital Humanities Quarterly 8 (2014), http://www.digitalhumanities.org/dhq/vol/8/1/000168/000168.html.

7 Cf. ibid.

8 https://voyant-tools.org/ [accessed: 14.05.2019].

9 http://www.laurenceanthony.net/software/antconc/ [accessed: 14.05.2019].

quite easily understandable interface.[10] This, in turn, has the deficiency that there is no opportunity to modify the software in relation to one's questions – at least not for anyone with a lack of suitable software skills. So quite the contrary becomes necessary, one's own questions have to be fitted to the software.[11]

In the following essay, I will (1) present the process of creating a digital corpus and possibilities to test the functionality of dirty OCR. Then (2), two different approaches of working with digital tools will be shown. First (2.1), data mining as a process of searching for trends and patterns without any strong presuppositions will be introduced. Second (2.2), the ability of dirty OCR and digital tools will be checked when (re)examining hypotheses gained by close or surface reading.[12] Essential questions will focus on the validity of produced data and what kind of research issues can be handled with dirty OCR and free available tools. In the end (3), there will be an answer to the question if it is worth investing time and work into digitization when the outcome is unavoidably provisional and erroneous.

1. The creation of a digital corpus

Pretty soon after deciding to give Digital Humanities a try it became clear to me that it would be impossible to digitize the whole corpus of four to five volumes of six different periodicals – comprising at least 30,000 pages.[13] So, a selection had to be made. Because much of the conception of the project is based on close and surface reading of *Süddeutsche Monatshefte*, [14] said peri-

10 For a further discussion of open tools cf. the contribution of Helene Schlicht in this volume.

11 For a discussion of the "epistemological proposition" of any tool, cf. *Rieder, Bernhard/Röhle, Theo*, Digital Methods: Five Challenges, in: David M. Berry (ed.), Understanding Digital Humanities, New York: Palgrave Macmillan, 2012, 67–84, esp. 68–71.

12 Cf. for a plea for "surface reading" as a third mode of reading between "close" and "distant reading" *Collier, Patrick*, What is Modern Periodical Studies?, in: Journal of Modern Periodical Studies 6 (2015), 107–108.

13 The whole print corpus consists of the war year volumes of *Neue Rundschau*, *Deutsche Rundschau*, *Die Zukunft*, *Die Gegenwart*, *Die Gesellschaft*, and *Süddeutsche Monatshefte*.

14 For an overview of some of the main hypotheses of the project cf. *Kauffmann, Kai*, Wissensvermittlung, Kulturpolitik und Kriegspropaganda: Thesen zur Kriegspublizistik der

odical was chosen. The plan was to not only validate the results of the digital tools through a comparison with the results of a human reader but also, vice versa, to check the results and main hypotheses of the human reader against those of the reading machine. Despite this limitation, there still remained around 7,000 pages to be digitized.

All these pages had to be scanned manually. Everyone who ever made scans for reading on screen knows that it is not that bad when pages are slightly crooked but it might lead to serious problems when those scans are meant to be further processed for the means of digital tools. The issues of *Süddeutsche Monatshefte* that were accessible are bound book-like volumes with up to 1,000 pages each. Especially in the middle of such big books usual scanners struggle with the book fold which can produce slanted or blurry pictures. So we not only needed to scan every single page but also had to review every digitized page, and, if necessary, scan it again or straighten it digitally, which required additional hours of work.

The scanned PDF-files were then transformed with OCR, using *Tesseract*, into raw text files with no further mark-ups and annotations.[15] Figures 1a and 1b show an original scan alongside the OCR transformed .txt-version that served as a basis for the work with digital tools.[16] Obviously, this is far off the demands of editorial philology and would be wholly inadequate as a basis for a digital or non-digital edition in the proper sense.[17] To make the results suitable for that, it would actually be necessary to invest even more time to correct the errors, which would still have to be conducted manually for the most part, despite all progress in training software to reach better results. But as time and resources were limited, that could not be done in this project.

---

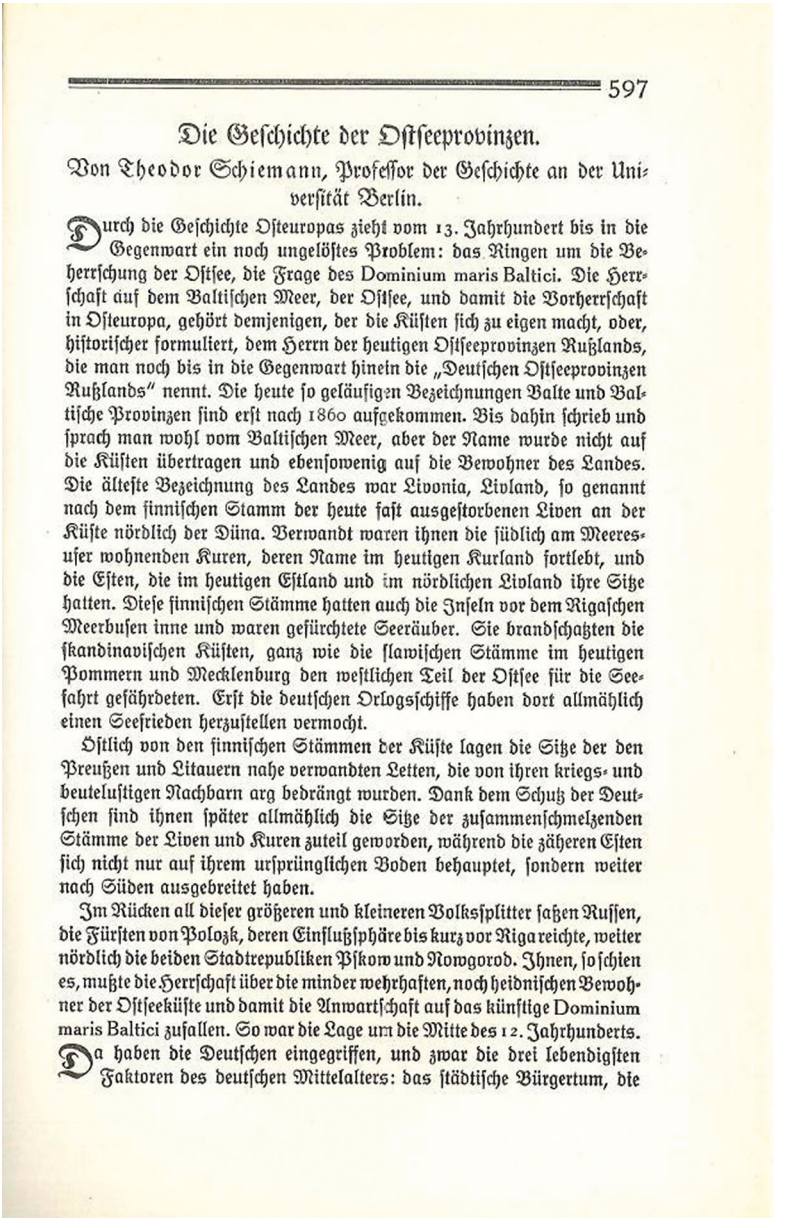
deutschen Rundschauzeitschriften 1914–1918, in: Olivier Agard/Barbara Beßlich (eds.), Krieg für die Kultur? Une guerre pour la civilisation? Intellektuelle Legitimationsversuche des Ersten Weltkriegs in Deutschland und Frankreich (1914–1918), Berlin: Peter Lang, 2018, 113–128.

15 For an expert description of this process, see the article of Patrick Jentsch and Stephan Porada in this volume.

16 For methods of measuring OCR accuracy that were not applied here cf. *R. Holley*, How Good Can It Get?.

17 At the present moment, the standard for digital editions is defined by the guidelines of the Text Encoding Initiative (TEI); cf. https://tei-c.org/guidelines/ [accessed: 14.05.2019].

Fig. 1a: Original scan of a magazine page



597

Die Geschichte der Ostseeprovinzen.

Von Theodor Schiemann, Professor der Geschichte an der Universität Berlin.

Durch die Geschichte Osteuropas zieht vom 13. Jahrhundert bis in die Gegenwart ein noch ungelöstes Problem: das Ringen um die Beherrschung der Ostsee, die Frage des Dominium maris Baltici. Die Herrschaft auf dem Baltischen Meer, der Ostsee, und damit die Vorherrschaft in Osteuropa, gehört demjenigen, der die Küsten sich zu eigen macht, oder, historischer formuliert, dem Herrn der heutigen Ostseeprovinzen Rußlands, die man noch bis in die Gegenwart hinein die „Deutschen Ostseeprovinzen Rußlands" nennt. Die heute so geläufigen Bezeichnungen Balte und Baltische Provinzen sind erst nach 1860 aufgekommen. Bis dahin schrieb und sprach man wohl vom Baltischen Meer, aber der Name wurde nicht auf die Küsten übertragen und ebensowenig auf die Bewohner des Landes. Die älteste Bezeichnung des Landes war Livonia, Livland, so genannt nach dem finnischen Stamm der heute fast ausgestorbenen Liven an der Küste nördlich der Düna. Verwandt waren ihnen die südlich am Meeresufer wohnenden Kuren, deren Name im heutigen Kurland fortlebt, und die Esten, die im heutigen Estland und im nördlichen Livland ihre Sitze hatten. Diese finnischen Stämme hatten auch die Inseln vor dem Rigaschen Meerbusen inne und waren gefürchtete Seeräuber. Sie brandschatzten die skandinavischen Küsten, ganz wie die slawischen Stämme im heutigen Pommern und Mecklenburg den westlichen Teil der Ostsee für die Seefahrt gefährdeten. Erst die deutschen Orlogsschiffe haben dort allmählich einen Seefrieden herzustellen vermocht.

Östlich von den finnischen Stämmen der Küste lagen die Sitze der den Preußen und Litauern nahe verwandten Letten, die von ihren kriegs- und beutelustigen Nachbarn arg bedrängt wurden. Dank dem Schutz der Deutschen sind ihnen später allmählich die Sitze der zusammenschmelzenden Stämme der Liven und Kuren zuteil geworden, während die zäheren Esten sich nicht nur auf ihrem ursprünglichen Boden behauptet, sondern weiter nach Süden ausgebreitet haben.

Im Rücken all dieser größeren und kleineren Volkssplitter saßen Russen, die Fürsten von Polozk, deren Einflußsphäre bis kurz vor Riga reichte, weiter nördlich die beiden Stadtrepubliken Pskow und Nowgorod. Ihnen, so schien es, mußte die Herrschaft über die minder wehrhaften, noch heidnischen Bewohner der Ostseeküste und damit die Anwartschaft auf das künftige Dominium maris Baltici zufallen. So war die Lage um die Mitte des 12. Jahrhunderts. Da haben die Deutschen eingegriffen, und zwar die drei lebendigsten Faktoren des deutschen Mittelalters: das städtische Bürgertum, die

Fig. 1b: The OCR transformed .txt-version of the same page

597
Die Geschichte der Ostseeprovinzen
Von Theodor Schiemann, Professor der Geschichte an der Uni-
Versität Berlin.
Durch die Geschichte Osteuropas zieht vom 13. Jahrhundert bis in die Gegenwart ein noch ungelöstes Problem: dasNingen um die Be-
herrschung der Ostsee, die Frage des Dominium maris Baltici. Die Herr-
schaft aus dem Baltischen Meer, der Ostsee, und damit die Borherrschaft
in Osteuropa, gehört demjenigen, der die Küsten sich zu eigen macht, oder, historischer formuliert, dem Herrn der heutigen Ostseeprovinzen Nußlands, die man noch bis in die Gegenwart hinein die „Deutschen Ostseeprovinzen Rußlands« nennt. Die heute so geläufigen Bezeichnungen Balte und Bal-tische Provinzen sind erst nach 1860 aufgekommen. Bis dahin schrieb und sprach man wohl vom Baltischen Meer, aber der Name wurde nicht auf
die Küsten übertragen und ebensowenig auf die Bewohner des Landes.
Die älteste Bezeichnung des Landes war Livonia, Livland, so genannt
nach dem finnischen Stamm der heute fast ausgestorbenen Liven an der
- Küste nördlich der Düna. Verwandt waren ihnen die südlich am Meeres-
ufer wohnenden Kuren, deren Name im heutigen Kurland fortlebt, und
die Esten, die im heutigen Estland und im nördlichen Livland ihre Sitze
hatten. Diese finnischen Stämme hatten auch die Jnseln vor dem Nigaschen Meerbusen inne und waren gesürchtete Seeräuber· Sie brandschatzten die skandinavischen Küsten, ganz wie die slawischen Stämme im heutigen Pommern und Mecklenburg den westlichen Teil der Ostsee für die See-
fahrt gefährdeten. Erst die deutschen Orlogsschiffe haben dort allmählich einen Seesrieden herzustellen vermocht-
Ostlich von den finnischen Stämmen der Küste lagen die Sitze der den Preußen und Litauern nahe verwandten Letten, die von ihren kriegs- und beutelustigen Nachbarn arg bedrängt wurden. Dank dem Schutz der Deut-schen sind ihnen später allmählich die Sitze der zusammenschmelzenden Stämme der Liven und Kuren zuteil geworden, während die zäheren Esten sich nicht nur aus ihrem ursprünglichen Boden behauptet, sondern weiter nach Süden ausgebreitet haben.
Jm Rücken all dieser größeren und kleineren Volkssplitter saßen Russen,
die Fürsten von Polozk, deren Einflußsphäreb is kurz vor Nigareichte, weiter nördlich die beiden Stadtrepubliken Pskow und Nowgorod. Ihnen, so schien es, mußte die Herrschaft über die minder wehrhaften, noch heidnischen Bewoh-ner der Ostseeküste und damit die Anwartschaft aus das künftige Dominium maris Baltici zufallen. So war die Lage um die Mitte des 12. Jahrhunderts·
Oa haben die Deutschen eingegriffen, und zwar die drei lebendigsten Faktoren des deutschen Mittelalters: das städtische Bürgertum, die

Fortunately, many of the OCR-induced errors are recurring and can thus be taken into account in the use of text mining tools. Among the most common errors – typically in the OCR of Fraktur fonts – are the transformation of the capital "R" to a capital "N", the capital "I" to a capital "J" and the lower case "s" to a lower case "f" or vice versa – though these transformation errors do not occur anytime.[18] So if you want to find results including the term "Rußland"

18 Typical errors further involve umlauts and end-of-line hyphenation; cf. figure 1b with the highlighting of some typical errors; cf. also *Riddell, Allen Beye*, How to Read 22, 198 Journal

(Russia), for example, it is no problem to search for "Rußland" or the incorrect "Nußland".[19]

Since text mining tools are not only valuable because they provide the possibility to search for single terms but also because they can identify structures, patterns and trends not yet recognized, there was need for closer scrutiny of the potential of the dirty OCR files. I worked with Voyant Tools for the most part, a web-based open-source software for computer philologist text analysis that offers a variety of tools and visualization.[20] As a basis for testing, the whole corpus was split up into fifty files representing the fifty scanned issues of *Süddeutsche Monatshefte* to find out if Voyant was able to identify the regional focus of them.[21] It was first browsed and then searched for occurrences of "Schweiz" (Switzerland). The results were striking and revealed a peak for the May 1916 issue, which mainly focuses on "Die Schweiz im Krieg" ("Switzerland at War") and is the only issue with a focus on Switzerland in the corpus. Similar results with similar preconditions were reached when browsing for "Vatikan" (Vatican) and "Spanien" (Spain). When searching for "England" and "Frankreich" (France), two of Germany's main enemies during World War I, things got a bit more blurry due to the significant rise of results. But comparing the tools results with the printed tables

Articles: Studying the History of German Studies with Topic Models, in: Lynne Tatlock/Matt Erlin (eds.), Distant Readings: Topologies of German Culture in the long nineteenth Century, New York: Rochester, 2014, 95.

19 It proved to be irrelevant whether both words were searched separately or together in form of a regular expression.

20 See footnote 7. The decision fell for Voyant instead of AntConc – which provide the same tools to a large extent –, on the one hand because of its more extensive visualization capabilities and, on the other hand and particularly, because of its integrated stop word list that greatly facilitates its use.

21 Since the beginning of the war, every issue of *Süddeutsche Monatshefte* had a main topic, ranging from domestic affairs to economic problems or geographical regions. For a schematic table of contents cf. *K. Kauffmann*, Wissensvermittlung, Kulturpolitik und Kriegspropaganda, 121–22. There would have been other ways of testing the functionality of the dirty OCR, of course, including automatized techniques. Anyway, the latter would have required a 'clean' subcorpus for comparison. So the decision fell for checking it manually. Searching for country names turned out to be a good way, firstly, because of the focus of the research project on comparisons between cultures based on nation states and, secondly, because of the availability of the tables of contents which often reveal the regional focus in their header.

of contents showed that Voyant was still able to identify those issues in the corpus with a focus on England and France.

Now how about the search for occurrences of "Russia" with the above-mentioned complications in the transformation into OCR readable text? Figure 2 is a visualization of the raw frequencies of "Rußland" and "Nußland" in the whole corpus. It shows significant peaks for the February, July and December 1915, March 1916 and January 1917 issues. Using the tables of contents as a basis for testing again, the results are mixed. On the one hand, Voyant identifies some of the special issues on Russia[22] and Russia as a thematic priority in other issues;[23] on the other hand, Voyant shows the highest peak for occurrences of "Rußland" and "Nußland" for the January 1917 issue. Though the guiding theme of this issue is "Äußere Politik" (foreign policy), judging by their titles, none of the articles seems to focus on Russia. A closer look at the text with the guiding help of the reader function of Voyant and the visualization of the distribution of "Rußland" and "Nußland" in the text can reveal the reason for this surprising insight. It is owed to an article by Graf Ernst Reventlow on the Turkish straits and their development.[24] Of course, anyone with enough knowledge about the Ottoman Empire or Turkey in the 19th century and during World War I could suppose the importance of Russia in such an article when finding it in the table of contents; nevertheless, this example shows the potential of digital tools to reveal what might remain hidden to a cursory look.

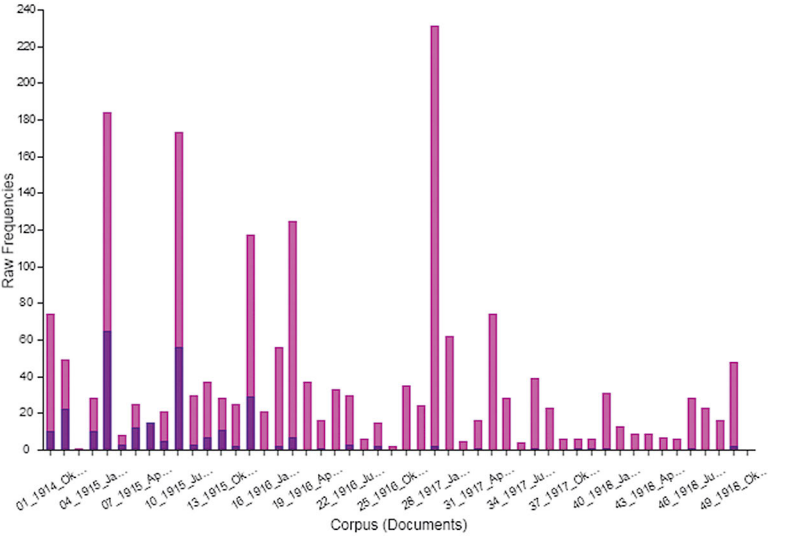
The most irritating outcome was undoubtedly the result for the October 1918 issue on the first anniversary of the Bolshevik Revolution in Russia. What was expected to be one of the highest peaks in the visualization is in fact only of medium height. At first, I was apt to blame it on dirty OCR. But that was wrong. Instead, the document length is responsible for the outcome. As Voyant shows in its summary section, the October 1918 issue belongs to the five shortest issues in the whole corpus – probably due to paper shortage

22 The February 1915 issue on "Rußland" ("Russia") and the July 1915 issue on "Rußland von Innen" ("Russia from the Inside").

23 This applies to the December 1915 issue, whose main topic are "Kriegsziele" ("War Aims") and which contains at least some articles on Russia, and to the March 1916 issue, whose title is "Kriegsgefangen"("War Captivity") and which features many articles on Russia.

24 *Reventlow, Ernst*, Die Frage der türkischen Meerengen und ihre Entwicklung, in: Süddeutsche Monatshefte 14 (1917): 432–66.

Fig. 2: A visualization of the raw frequencies of "Rußland" and "Nußland" in the whole corpus



towards the end of the war – with an amount of 35,463 (recognized) tokens – in contrast, the longest issue has an amount of 90,518 (recognized) tokens.

So this and the previous example do not reveal that much about the quality of the OCR and its usability but about the formal structure and the contents of the periodical it is based on. First, there are the above-mentioned differences in the length of each document representing a single issue that produce results different from those we would expect with the tables of content in mind.[25] In this case, the described results are no huge surprise, for sure, but they could have occurred with other words and in other contexts.[26]

25 A possible solution to this problem might be the calculation of the median value; cf. *Jannidis, Fotis/Kohle, Hubertus/Rehbein, Malte (eds.)*, Digital Humanities: Eine Einführung, Stuttgart: Springer, 2017, 282–83. Nevertheless, the inexperience of most traditionally educated humanists with statistics and mathematics comes into play here. Anyone engaging deeper in Digital Humanities has to learn whole new things to check the data for its reliability.

26 See, for example, *Roff, Sandra*, From the Field: A Case Study in Using Historical Periodical Databases to Revise Previous Research, in: American Periodicals 18 (2008), 96–100. In her brief account, she emphasizes the great opportunities of keyword searches in

Second, a single article within the issues of an outstanding length might lead to some distortions in the results. This is the case with the above-mentioned article on the Turkish straits and is more often the case in *Süddeutsche Monatshefte* and similar magazines.[27]

The OCR itself is certainly not good enough to identify each and every token of a given word. But considering the fact that Digital Humanities is used for the recognition of larger structures and patterns, too, this is not necessary at all. The dirty OCR seems capable of identifying such trends and might thus be suited to serve as an instrument at least for testing hypotheses developed through close readings of selected parts or surface readings of larger parts of the corpus in question.

2. Working with dirty OCR

Ultimately, all tools for text analysis in Digital Humanities serve two different purposes. On the one hand, they can search Big Data for patterns and structures which are not discovered yet or are not to be discovered at all with the means of traditional reading techniques – this is generally known as "data mining". On the other hand, insights gained by traditional reading techniques can be tested with the means of Digital Humanities.[28] Though both approaches can be combined, it is obvious that both of them require different forms of engagement. While one can simply use the abilities of tools to reveal frequencies or patterns for further investigation in the first case, it is inevitable to consider beforehand what one actually wants to find out in the second case. Both options will be tested in the following sections for their hermeneutical significance as well as for their potential with a corpus full of OCR noise.

---

big databases, where previously only titles of articles served as hints for further investigation.

27 There is a significant difference between newspapers and book-like journals: The former are not as prone to distortions because of their manifold content and the smaller extent of their articles.

28 In a slightly different context this is the differentiation between a "corpus-driven approach" and a "corpus-based approach"; cf. *Anthony, Laurence*, A Critical look at software tools in corpus linguistics, in: Linguistic Research 30 (2013), 141–61.

2.1 Data mining with dirty OCR

The text corpus was uploaded to Voyant without any problems. One of the first things it provides is a word cloud which displays the most frequent words.[29] As figure 3 shows, dirty OCR definitely causes problems.

Fig. 3: A word cloud without additions to the stop list …

This becomes most obvious when looking at single characters like "i", "a" and "e" and at the German prefix "ge-", probably split off from words because of bad character recognition or problems with end-of-line hyphenation. Further, there are words like "jn" and "jch", correctly spelled with an "i". The words "in" (in) and "ich" (I) are usually detected by the integrated stop list of Voyant, which, to a large extent, consists of function words with less hermeneutical significance. Due to errors in the character recognition, which transformed the Fracture "i" into a "j", the stop list does not work correctly either. This is undoubtedly annoying but still a manageable problem. It is quite easy to adjust the stop list by adding single characters, prefixes and numbers. After making additions to the stop list for each error and five iter-

29 It is only the font size and position in the cloud that reveals the importance of the words due to their frequency; the different colors of the words produced by Voyant are of no significance and thus not depicted here.

ations, I arrived at the word cloud shown in figure 4. This cloud is free from OCR noise and could serve as a starting point for interpretation.[30]

Fig. 4: … and with additions to the stop list

Apparently, there are some commonalities between both word clouds. On top, there is the central position and large point size of words with the stem "deutsch-" (German).[31] What does that mean? Well, this is the point where text mining stops and interpretation begins because no tool will ever tell anything about the hermeneutical significance of its outcome. In this case, the results might not be very surprising at first sight. Due to the fact that the

30 For a plea for digital corpus linguistics including especially the counting of word frequencies as a starting point for further investigations see *Archer, Dawn*, Data Mining and Word Frequency Analysis, in: Gabriele Griffin/Matt Hayler (eds.), Research Methods for Reading Digital Data in the Digital Humanities, Edinburgh: Edinburgh University Press, 2016, 72–92.

31 Another problem becomes obvious, here. It is the lack of any further lemmatization in the raw text files as well as in Voyant's stop list. This means, that any deviant grammatical form of a given word is recognized as a new type of word. To some extent, this problem could be solved with the use of regular expressions though this is quite time consuming as well.

corpus consists of issues of a German journal printed during World War I which mainly addresses the German situation during the war, contains discussions of reasons for the war and of war aims as well as a justification for German warfare to a general public and to the enemies, it seems all too clear that "deutsch-" has such a statistical significance within the documents.

However, Kai Kauffmann claims that the military conflict led to an increase of international coverage of the nations involved in the war.[32] Indeed, he has good reasons for his assertion, since the guiding themes of more than half the issues of *Süddeutsche Monatshefte* focus on foreign nations or regions. One reason for this apparent contradiction might be the different length of the documents. But when I uploaded only those issues dealing with other nations than Germany to *Voyant*, the outcome was not substantially different with regard to the statistical importance of "German" (see figure 5). This finding is far away from refuting Kauffmann's assumption. Quite the contrary, the word cloud serves as a vehicle to refine it.

Fig. 5: Word cloud of only those issues dealing with other nations than Germany

32 *K. Kauffmann*, Wissensvermittlung, Kulturpolitik und Kriegspropaganda, 120: "Allerdings führte der militärische Konflikt zu einer Verstärkung der internationalen Berichterstattung über die am Krieg beteiligten Staaten." This aspect is discussed again further below in this article in section 2.2 with another focus.

In a project concerned with publicists comparing nations and cultures, the result can be interpreted as that the growing interest in other nations during World War I does not constitute an actual interest in the nations themselves. Instead, every statement and opinion piece about other countries contained in the journal seems to only be relevant in relation to Germany. The nature of these relationships is, however, not visible in the word cloud but needs to be further investigated by means of other tools[33] or through close readings of single articles and issues.[34]

There are some more commonalities between the word clouds in figure 4 and 5, in particular the visible significance of the main German war opponents England, France and Russia[35] – and the United States? It is surprising that they are not part of this cloud, even in view of the fact of their late entry into the war in 1917. At least there are two issues of *Süddeutsche Monatshefte* that have a focus on the USA and some more articles in other issues. Are they missing because of dirty OCR? I doubted that because in general the character recognition worked quite well; instead, at some point, I doubted the functionality of Voyant and presumed it stopped processing the corpus somewhere in the middle.[36] This assumption was disproved, however, when I finally found the reason, namely the lack of any further annotation, espe-

33 Some possibilities for similar problems will be discussed in section 2.2 below.

34 Though it is likely, that comparisons are quite important within the constructed and presented relations, the reasons for such a strong orientation towards similarities and distinctions across nations are not so evident. Two main aspects might be at work: On the hand, the situation of war might play a role. Measures and means of administrating occupied territories have to be discussed as well as strengths and weaknesses of the enemies to assess the possibility of victory or probable risks that could undermine victory. On the other hand, a more general aspect might be at work. Comparing other nations and/or cultures with one's own can serve as a means to arouse interest or to help understand the 'other'.

35 "England", "Frankreich", "Rußland".

36 In fact, Voyant might have problems like that in some cases due to server capacities and limited working memory – a problem that arose in my work with Voyant when trying out its tool for topic modeling. More general, this leads to the question of epistemic trust in the tools' functionality: As long as one is not able to read the codes of the tool, there remains nothing more than to trust in the produced data. However, there is a quite simple solution that might help at least in the case of large corpora where it is not possible to check the results with close reading: the use of "different tools from the same category" – cf. *B. Rieder/T. Röhle*, Digital Methods, 77.

cially Named Entity Recognition (NER) and lemmatization. A closer look at the word cloud reveals the inconspicuous word "staaten" (states). What could simply be the plural form of "Staat" (state) – and in some cases stands for nothing more – turns out to be a part of the term "vereinigte staaten" (united states) in many instances.[37] Without annotation, Voyant treats "Vereinigte Staaten" (United States) not as a single term but splits it into two terms. Moreover, when testing for instances of "amerika" (America), which in most cases is used synonymously with "United States", the term turned out to appear only 701 times in the corpus. In comparison, "frankreich" (France) is used 1976 times. Then again, "amerika*" can be found 2,074 times in the document. This clearly shows that digital tools without annotated corpora do have their limits when processing inflectional languages like German. It also demonstrates the need to pay close attention to the functioning of any tool and to the condition of any digitized corpus. This is especially true when one compares prior knowledge – or rather expectations – to actual results. Great differences between them do not necessarily rest upon wrong expectations but could be the result of a malfunctioning tool or somehow flawed documents.

Let us turn from here to the differences between figure 4 and 5. One of them is visible at the bottom left of figure 5. "belgien" (Belgium) is missing in the word cloud displaying all the issues and comes into play only in the word cloud displaying the issues on other countries than Germany. Moreover, it is the only smaller nation involved in World War I which is shown[38] – while even Germany's most important ally, Austria, is missing. This intriguing result leaves room for speculation. Was Belgium more important to the authors of *Süddeutsche Monatshefte* than Austria because it was German-occupied and they therefore felt the need to discuss means of administration in the Belgian territory? Or was it because allied reports on German atrocities in Belgium had to be denied?[39]

37 This was tested with the help of Voyant's contexts tool – but any tool with the possibility to show the amount of counted words would work as well – with the result of 426 instances of "vereinigte staaten" and "vereinigten staaten" and a total number of 1,150 instances of "staaten".

38 The word would have been depicted even bigger if Voyant had integrated instances of "velgien", which stems from bad OCR.

39 For the international discussion of German war atrocities in Belgium cf. the extensive study of *Horne, John/Kramer, Alan*: German Atrocities, 1914: A History of Denial, New Haven: Yale University Press, 2002.

In fact, the result is misleading, at least partially. Using Voyant's document terms tool to get an overview of word occurrences reveals that instances of "*sterreich"[40] (836 instances) exceed instances of "*elgium" (699) in the whole text corpus. Only in the subcorpus containing issues on other nations than Germany does the use of "*elgium" (631) exceed instances of "*sterreich" (529). It can be concluded that overall Belgium is not considered more important than Austria. Instead, this comparison reveals a tendency to write articles on Austria not so much in contexts of foreign nations but in relation to the writers' own German nation.

Nevertheless, the speculations about the importance of Belgium to German writers still stand. The depicted word cloud provides no indication of what the articles are about, however. To shed light on this matter, I cut every article on Belgium out of the whole text corpus,[41] uploaded the collection to Voyant and arrived at the result displayed in figure 6.

Fig. 6: Word cloud of any article on Belgium

40 Searching for "österreich" (Austria) turned out to be not the best option because of the initial umlaut that leads to OCR errors in many instances.

41 This step was necessary partly because issues on Belgium had more central topics – for example, the April 1915 issue with another focus on Bismarck at the occasion of his hundredth birthday –, partly because articles from other issues could be taken into account. The selection was done by the guidance of the table of contents so that any article with a mention of Belgium in its title was included in this subcorpus.

What strikes the attention in this word cloud are – besides the expectable central position and font size of words with the stem "belg*", once again, the high frequency of words with the stem "deutsch*" (German), the occurrence of France and England as Germany's main enemies at its western front, and the terms "krieg" (war) and "neutralität" (neutrality) – essentially two things. The first is a concentration of urban spaces like "brüssel" (Brussels), "antwerpen" (Antwerp) and "küste" (coast). While this clearly stems from the requirements and conditions of the war, there is furthermore a focus on the main ethnic groups in Belgium, namely "flamen" or "vlamen" (Flemings) and "wallonen" (Walloons). In fact, this gives a hint for the representation of a central German perspective on Belgium during World War I in *Süddeutsche Monatshefte*. In political as well as in media debates there was a concentration on the Flemings as a seemingly Germanic people which had to be protected from French or rather Romanic influence and oppression in a multilingual state. Underlying concepts were based on ideas of divide and conquer, the fear of encirclement, and ultimately *völkische* and racist notions of nations and their structures.[42] Of course, this cannot be derived directly from the word cloud. Rather, it is based on prior knowledge originating from close readings of articles in other journals and from readings of research on Belgium during World War I. So it is unlikely but possible that the articles in *Süddeutsche Monatshefte* with their focus on Flemings and Walloons take a critical stance on this separation.

What helps, then, is the use of Voyant's contexts tool[43] to examine the use of the single tokens. It is not absolutely necessary to read every single instance in this tools panel – in this case, for example, some hundred appearances of "wallon*", "flam*", "fläm*",[44] and so on would have been to checked. Yet even a cursory look at the results confirms the above suppo-

42 Cf. *Schaepdrijver, Sophie De*, Belgium, in: John Horne (ed.), A Companion to World War I., Chichester: Wiley-Blackwell, 2010, 391–393. For a more detailed presentation of German images of Belgium during World War I cf. *Bischoff, Sebastian*, Kriegsziel Belgien: Annexionsdebatten und nationale Feindbilder in der deutschen Öffentlichkeit, 1914–1918, Münster/New York: Waxmann, 2018.

43 Tools like this are usually known as Keyword in Context (KWIC). They generate a list of any instance of a word in question with a variable context of words on the left and right side of the given token.

44 Here, again, the lack of NER and tokenization is annoying for any instance of these words has to be searched separately.

sition: Indeed, a good part of the instances is concerned with a definite differentiation and separation of the Flemings and Walloons and with the attempt to highlight the Flemings as a Germanic people. At this point, the term "sprache" (language) in the word cloud comes into play. When using the contexts tool again, the results highlight the significant connection of language with the Flemings in the documents, whereby Flemish is presented as a suppressed language, which has to be protected and supported by the Germans.

Now this might lead to the presumption that instances of the aforementioned terms occur in the context of terms like "verwaltung" (administration), "regierung" (government, administration), and "politik" (politics, policy) and that there might be proposals in the articles for an occupational administration to the advantage of the Flemings. However, this is only correct in some way. In fact, the occurrences of "politik" refer to the foreign politics of the former Belgian government and the British government in most cases; and the occurrences of "regierung" refer to the domestic and foreign politics of the former Belgian government most often. Only the term "verwaltung" is used in the supposed sense. Does that mean that in the majority of cases the contributors of *Süddeutsche Monatshefte* argue in favor of administrational measures under military occupation of Belgium instead of an annexation? At least this is what Kauffmann supposes based on his close reading of the April 1915 issue on Belgium.[45]

But there is another problem with the term "verwaltung". As in the example of the article on the Turkish straits mentioned above, a further examination of the results reveals that its appearance in the word cloud is bound to a good part to one single article on Belgium under German administration.[46] This does not necessarily mean that governmental and administrational questions in occupied Belgium are of minor importance than the word

45 Cf. *K. Kauffmann*, Wissensvermittlung, Kulturpolitik und Kriegspropaganda, 125.

46 *Bissing, Friedrich Wilhelm Freiherr von*, Belgien unter deutscher Verwaltung, in: Süddeutsche Monatshefte 12 (1915), 74–93. The same is true for the word "unterricht" (education, teaching) whose occurrence is bound to a good part to the article of *Ziegesar, Jozef Haller van*, Der mittlere Unterricht in Belgien, in: Süddeutsche Monatshefte 13 (1916), 605–616. This also points to a yet unmentioned problem with the underlying raw text files. Since there is no further annotation, page headers with the name of the author and/or the title of the article are always counted as a new token of a word and thus lead to erroneous results.

cloud – and Kauffmann in his article – suggest; then again, these topics probably are to be recognized on another level than words. One possibility could be the connection and arrangement of articles in the whole issue as discussed below.

2.2 Testing for hypotheses with dirty OCR

The discussion in the section above was primarily led by a deductive method, using Voyant as a means to reveal some of the foci of *Süddeutsche Monatshefte* and to go deeper until a point is reached where close reading seems to be the best option for further investigation. For this approach, no or less presuppositions were needed. However, as mentioned above, there are some strong assumptions about the corpus developed and published by Kai Kauffmann. Is it possible to reassess them with dirty OCR and Voyant? In the following section, some of these assumptions will be presented and possibilities to test them will be discussed.

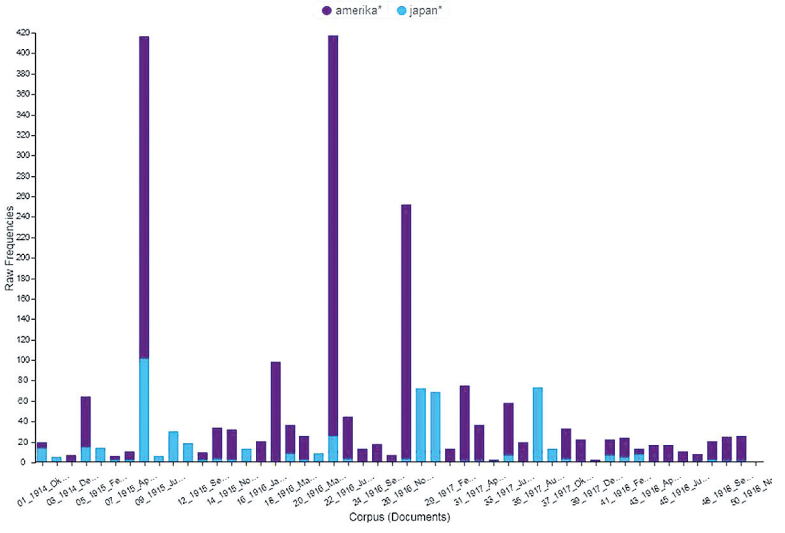
(1) Kauffmann suspects at least a small increase of globalized horizons of comparison in German wartime journalism especially due to the entry of the USA and Japan into the war.[47] This assumption seemed to be easy to prove. Using Voyant's trends tool, I searched the document for "amerika*" (America)[48] and "japan*"[49] with the result depicted in figure 7. The columns indeed show a continuous occurrence of both terms throughout the corpus with some significant peaks. When asking for the reason of those peaks, the solution was somewhat obvious though disappointing for any height rested upon articles and main topics, which could have been identified easily with a close look on the table of contents. In this light, the advantage of digital analysis tools is only a faster result compared to counting the articles manually.

47 *K. Kauffmann*, Wissensvermittlung, Kulturpolitik und Kriegspropaganda, 120.

48 The search for "USA" and "Vereinigte Staaten" (United States) turned out to be of not even minor significance; for the reasons cf. the discussion on single terms and the problems with missing annotations above in section 2.1.

49 In the face of dirty OCR, I searched for instances of "iapan*" as well. But as there were only two results in the whole corpus, I left those instances out of account.

Fig. 7: Raw frequencies of "amerika" and "japan*" in the whole corpus*



Nevertheless, the created visualization goes beyond Kauffmann's insight in some respect. On the one hand, it is a means for a more precise and striking presentation of what stays kind of vague in Kauffmann's article.[50] On the other hand, it reveals a possible decline of globalized comparisons at the turn of 1916/1917 – at least there is no more actual focus on Japan or the United States. This is especially surprising in light of the United States' entry into the war in April 1917, an event producing lots of media coverage, one would think. Is it the monthly publication frequency of the journal that makes it difficult to react to even such an important event? Or do events and circumstances at the home front become more important?[51] Further investigation in this respect

50 Moreover, it calls attention to the nearly always quite careless handling of quantifiable statements in the philologies. Of course, quantification in a statistical sense is no adequate option in any case – be it because an exact value is of no further explanatory power or be it because no digitized corpus is available. But in some instances, digital tools definitely help to underline insights which would otherwise be nothing more than unverifiable claims. On the contrary, statistical outcomes are far away from being self-explanatory. For a further discussion of these epistemological questions of Digital Humanities cf. *B. Rieder/T. Röhle*, Digital Methods, 71–79.

51 There are some hints for this supposition because many of the later wartime issues of *Süddeutsche Monatshefte* deal, for example, with German agriculture (July 1917), German

is needed, but the potential of digital tools to move from an initial question to a quite different one, which was out of sight before, becomes obvious.

(2) Another important finding of Kauffmann is the outstanding position of historians within the contributors of *Süddeutsche Monatshefte*. In contrast, philologists, or especially natural scientists, play only a marginal role, if at all.[52] Follow-up questions could aim at differences in the use of distinctive terms or patterns of arguments depending on the academic profession of the contributors. Once again, it is not dirty OCR that causes problems but the lack of any further annotation, in this case regarding metadata on the contributors and their academic profession – a gap that should be closed for further work with the digital corpus and its analysis with digital tools, let alone the publication of a digital edition of the documents –, which makes the examination difficult.[53] Of course, also in this case the selection and formation of a subcorpus could be done; but, in the end, this is a whole new stage of work consuming lots of time and therefore costs – work I have not done yet so I am unable to present any results.

(3) Finally, there is an essentially theoretical perspective on the *form* of periodicals in Kauffmann's approach, asking for their special ability to build up comparisons – or opportunities to compare – due to the arrangement of their material.[54] Almost any journal brings together articles by different authors with the same or different views on the same or related topics,[55] thus enabling the reader not only to accumulate the knowledge but also to compare between those views. So journals can stabilize or destabilize existing opinions or even formations of discourse. Here, the focus is not so much on

social democracy (November 1917), German industry (March 1918), or German workers (January 1918).

52 *K. Kauffmann*, Wissensvermittlung, Kulturpolitik und Kriegspropaganda, 127.

53 The same problem arises, by the way, when asking for differences between different factional and fictional genres. Even when testing for the ability of a tool to differentiate between, genres a comparative corpus generated by a human being is needed; cf., for example: *Allison, Sarah et al.*, Quantitive Formalism: an Experiment, accessed: 14.05.2019, https://litlab.stanford.edu/LiteraryLabPamphlet1.pdf.

54 *K. Kauffmann*, Wissensvermittlung, Kulturpolitik und Kriegspropaganda, 124–26.

55 Exceptions are, for example, journals that are not only edited but written to a large extent by only one person like Karl Kraus' "Die Fackel" or Maximilian Harden's "Die Zukunft". Other exceptions might be periodicals with a powerful editorial board and a rigorous political agenda.

the content of single articles but on the arrangement of an 'ensemble of texts'[56] and therefore on the media preconditions of practices of comparison, their effects on practices of comparison and the modalities of the production of knowledge. In fact, this is by the far most difficult aspect to test with Voyant. To explain the underlying concept in more detail and to present the problems with Digital Humanities in this case, I will focus on the July 1915 issue of *Süddeutsche Monatshefte* which has an emphasis on "Rußland von Innen" ("Russia from the inside"). First, I will show some of the results of Voyant when processing this issue to then compare them with central insights gained by close reading.

Again, I started with a word cloud whereby I deleted any instances of "deutsch-" (German) and "russ-" in addition to the OCR errors to reveal more of the things beyond the expectable (see figure 8). Though there are a lot of remarkable objects in this cloud to focus on,[57] I will concentrate on one aspect only: women and the question of gender.

In the historical context of the document, it seems evident that there is a close relationship between women – depicted on the upper left side of the cloud ("frauen") – and family-related words like "kinder" (children) and "sohn" (son). A good way to test supposed correlations with Voyant is the employment of its collocates tool which shows words found in proximity of a keyword. As figure 9[58] reveals, there is indeed a correlation between women

56 *K. Kauffmann*, Wissensvermittlung, Kulturpolitik und Kriegspropaganda, 124.

57 Especially the central position of the word "jüdischen" (Jewish) would need closer attention. While in this case it is obviously caused by a longer article on the situation of the Jewish proletariat in Russia, it is remarkable that Jews are the only ethnic or religious group that appears in figure 5, too. Is this focus on the Jews a result of antisemitism? Or is it because of the encounter with (orthodox) Jews in the occupied territories in the East? Do the contributors of *Süddeutsche Monatshefte* discuss possibilities to help the mainly poor and suppressed Jews in tsarist Russia or do they focus on measures to separate the Jews? For a general discussion of these questions cf., for example, *Zechlin, Egmont*, Die deutsche Politik und die Juden im Ersten Weltkrieg, Göttingen: Vandenhoeck u. Ruprecht, 1969; *Angress, Werner T.*, Das deutsche Militär und die Juden im Ersten Weltkrieg, in: Militärgeschichtliche Mitteilungen 19 (1976), 77–146; and *Hoffmann, Christhard*, Between Integration and Rejection: The Jewish Community in Germany, 1914–1918, in: John Horne (ed.), State, Society and Mobilization in Europe during the First World War, Cambridge: Cambridge University Press, 1997, 89–104.

58 The context is restricted to five words on each side; depicted are only those collocates which appear at least three times.

Fig. 8: Word cloud of the July 1915 issue of Süddeutsche Monatshefte *with some additions to the stop list*

and family, as "mutter" (mother) and "kinder" are among the top collocations. Not very surprising, too, is the occurrence of "männer" (men) in the list because talking about gender issues was – and still is – almost always talking about differences between the sexes and about their relationship. So what really draws attention are the words "sachalin" (Sakhalin), "revolution" and "beteiligung" (participation).

Fig. 9: Collocations of "frau" in the July 1915 issue of* Süddeutsche Monatshefte

Term	Collocate	Count (context)
frau*	revolution	5
frau*	männer	5
frau*	mutter	4
frau*	kinder	4
frau*	beteiligung	4
frau*	sachalin	3

The occurrence of "sachalin" is obviously due to the imprint of a report by Anton Čechov on the penal colony on the island Sakhalin.[59] But why does it feature women? The article includes a section on "Die Frauen and Kinder von Sachalin" ("The Women and Children of Sakhalin") where Čechov talks about women who committed a "crime, almost exclusively murder, which rests upon love affairs and family conflicts"[60] and women who followed their sentenced men to Sakhalin. Without any income opportunity, they would sooner or later engage in prostitution.[61] Yet this is the result of a close reading of the article while the digital tools did not reveal the exact same insights. Even if tools that help to reach comparable results exist, traditional forms of reading are the method of choice when it comes to analyzing a relatively short article like the one in question.

But what is revealed by this is the possible connection to two other articles of this issue: to an article by Adolf Dirr on "Die Russin"[62] ("The Female Russian") and to an article by Nadja Straßer on "Die russische Frau in der Revolution" ("The Russian Woman in the Revolution").[63] The titles of both texts already reveal their interest in gender issues but the constellation is remarkable in more than one respect. On the surface, a female and a male author simply voice their opinion on related topics in the two articles. While Dirr explicitly points to the fact that he is writing from a male standpoint,[64] there is an editor's note above Straßer's article which declares his contribution to represent the European view and hers to be the view of a "liberal Russian woman"[65] – by the way, a Jewish feminist who moved to Vienna in the late 1890s and lived in Berlin at the time of publication of

59 *Tschechow, Anton*, Die Gefängnisinsel Sachalin, in: Süddeutsche Monatshefte 12 (1915), 701–710.

60 Ibid., 708: "[…] Verbrechen, fast ausschließlich Mord, [die] auf Liebesaffären und Familienzwistigkeiten beruhen […]".

61 Ibid., 709.

62 *Dirr, Adolf*, Die Russin, in: Süddeutsche Monatshefte 12 (1915), 588–596.

63 *Straßer, Nadja*, Die russische Frau in der Revolution, in: Süddeutsche Monatshefte 12 (1915), 647–652.

64 Cf. *A. Dirr*, Die Russin, 588. For biographical information on Adolf Dirr cf. *Öhrig, Bruno*, Adolf Dirr (1867–1930): Ein Kaukasusforscher am Münchner Völkerkundemuseum, in: Münchner Beiträge zur Völkerkunde 6 (2000), 199–234.

65 *N. Straßer*, Die russische Frau in der Revolution, 647: "Wir nehmen an, daß es für unsere Leser Wert hat, nachdem sie in Dr. Dirrs Aufsatz den europäischen Maßstab angelegt

the article.[66] Thus cultural and gender aspects of the authors are interwoven, and the article is presented as an offer for comparison by the editor of *Süddeutsche Monatshefte* in light of different aspects of authorship.[67]

What do both address in detail? Dirr, while praising the Russian woman for "being more natural, less spoiled by culture" than the West European woman – thus connecting stereotypes about women and Russians as creatures of nature – focuses on her character, which he finds nevertheless to be "vain, empty, saucy, haughty, cheeky".[68] Straßer, in contrast, emphasizes the "spontaneity and certainty" ("Ungezwungenheit und Sicherheit") of the Russian woman that makes her tend to a comradely relationship to men and to revolutionary movements.[69] Most instances of the word "beteiligung" (participation) in the results of the collocates tool can be found in this text: It is the participation of women in the revolutionary action in 1905 in Russia.

Now, what might that mean for readers of *Süddeutsche Monatshefte* – predominantly male members of the educated bourgeoisie? Though Straßer's article is full of sympathy for female and social insurrection, it is framed by Dirr's article and Čechov's report. In this regard, when it comes to comparing the three texts, Straßer's perspective might be nothing more for contemporary readers than an affirmation of Dirr's chauvinistic view of Russian women, who, finally, end up in Čechov's penal colony for their tendency to insubordination. This is clearly not the result of the intention of the contributors or the message of the single articles. Instead, it is due to the compilation and arrangement of these articles in the same issue of *Süddeutsche Monatshefte*.

---

sahen, nun auch den spezifisch russischen Standpunkt vertreten zu sehen, indem wir einer freiheitlichen Russin das Wort geben."

66 Cf. *Schmidt, Birgit*, 'Die Frauenpflichtlerin' – Zur Erinnerung an Nadja Strasser, in: Aschkenas 16 (2006), 229–259.

67 The editor at this time was Paul Nikolaus Cossmann; cf. *Selig, Wolfram*, Paul Nikolaus Cossmann und die Süddeutschen Monatshefte von 1914–1918: Ein Beitrag zur Geschichte der nationalen Publizistik im Ersten Weltkrieg, Osnabrück: A. Fromm, 1967.

68 Dirr, "Die Russin," 592: "Eitel, leer, naseweis, hochmütig, vorlaut […]."

69 Straßer, "Die russische Frau in der Revolution," 649. Nevertheless, she reproduces stereotypical representations of Russians as well when calling them humans of emotion instead of action ("nicht Tat-, sondern Gefühlsmensch") and "half wild and primitive" ("halbwild und primitiv"); ibid., 647.

It is this potential of journals to produce meanings beyond single articles by addition of knowledge and claims and by comparison of knowledge and claims what makes them special and interesting for research. But these procedures of writing, editing, and reading rest upon structures that are hard to be detected with tools like those made available by Voyant – and, perhaps, rest upon structures so closely connected with human understanding that it is inevitable to return to close reading, even if one has started with distant reading.[70]

3. Conclusion

Without any doubt, dirty OCR is not appropriate for the production of robust and final results of any research. Too many errors make it impossible to detect any occurrence of certain words. Moreover, the lack of any further annotations, lemmatization or named entity recognition disturbs the quantifiable statistical outcome. Nevertheless, it works well enough when it comes to prove hypotheses about larger trends in a given corpus or to reveal something more about larger patterns. Therein lies the greatest potential of dirty OCR. Of course, when working with digital tools some adjustments have to be made – an addition of the stop list, for example, – and greater attentiveness to the (formal) structure of the digitized corpus is needed in contrast to a 'clean' corpus. But these problems are manageable to a great extent.

This is true, too, for the work with a web-based tool like Voyant. Publications like those of Franco Moretti and others have shown much more of the potentials of Digital Humanities, but their authors are conducting their research in a nearly perfect scientific environment with computer scientists who are available any time to refine the abilities of tools again and yet again. For those of us who lack this close connection with computer experts, free accessible tools – alongside clean or dirty OCR – are at least an option for starting research.

70 Cf. for example Stefan Scherer and Claudia Stockinger, "Archive in Serie: Kulturzeitschriften des 19. Jahrhunderts," in *Archiv/Fiktionen: Verfahren des Archivierens in Literatur und Kultur des langen 19. Jahrhunderts*, eds. Daniela Gretz and Nicolas Pethes (Freiburg, Berlin, and Wien: Rombach, 2016), 268, with their claim that digital tools can not reproduce the seriality of periodicals so that close reading is irreplaceable in periodical studies.

Whether dirty OCR and web-based tools can be of help in everyday research or not, largely depends on the research questions, of course. Voyant proved to be a valuable tool, especially when working with it in a mere deductive way with no or less presuppositions about the corpus. On the one hand, Voyant's word clouds made it was possible to go deeper and deeper into the corpus while creating new questions or refining preliminary assumptions. On the other hand, as the discussion above has shown, the effectiveness of Voyant and probably of any digital tool drops when the research focuses on a smaller number of articles.

In this regard, the production of documents with dirty OCR is probably most suitable for mid-scale corpora with a good rate of production time and possible research results – for the larger a corpus is, the more time is needed for its digitization. This time might be wasted when the digitization is not finished in accordance with (digital) editorial standards; the smaller a corpus is, on the contrary, the less will be the advantage of using digital tools.

In an academic world that is driven more and more to be project-based, which means the probable outcome of research should at best be known before any work has started, provisional digitized documents could definitely serve as a basis for the verification and/or development of hypotheses. Though not necessarily adoptable for publication, the results might at least be a basis for better-grounded assumptions in project proposals.

Bibliography

Allison, Sarah et al., Quantitive Formalism: an Experiment, accessed: 14.05.2019, https://litlab.stanford.edu/LiteraryLabPamphlet1.pdf.

Angress, Werner T., Das deutsche Militär und die Juden im Ersten Weltkrieg, in: Militärgeschichtliche Mitteilungen 19 (1976), 77–146.

Anthony, Laurence, A Critical look at software tools in corpus linguistics, in: Linguistic Research 30 (2013), 141–61.

Archer, Dawn, Data Mining and Word Frequency Analysis, in: Gabriele Griffin/Matt Hayler (eds.), Research Methods for Reading Digital Data in the Digital Humanities, Edinburgh: Edinburgh University Press, 2016, 72–92.

Bischoff, Sebastian, Kriegsziel Belgien: Annexionsdebatten und nationale Feindbilder in der deutschen Öffentlichkeit, 1914–1918, Münster/New York: Waxmann, 2018.

Bissing, Friedrich Wilhelm Freiherr von, Belgien unter deutscher Verwaltung, in: Süddeutsche Monatshefte 12 (1915), 74–93.

Collier, Patrick, What is Modern Periodical Studies?, in: Journal of Modern Periodical Studies 6 (2015), 107–108.

Cordell, Ryan, 'Q i-jtb the Raven': Taking Dirty OCR Seriously, in: Book History 20(2017), 194–95.

Dirr, Adolf, Die Russin, in: Süddeutsche Monatshefte 12 (1915), 588–596.

Hahn, Carolin, Forschung benötigt Infrastrukturen: Gegenwärtige Herausforderungen literaturwissenschaftlicher Netzwerkanalysen, in: Toni Bernhart et al. (ed.), Quantitative Ansätze in den Literatur- und Geisteswissenschaften: Systematische und historische Perspektiven, Berlin and Boston: de Gruyter, 2018, 315–34.

Hoffmann, Christhard, Between Integration and Rejection: The Jewish Community in Germany, 1914–1918, in: John Horne (ed.), State, Society and Mobilization in Europe during the First World War, Cambridge: Cambridge University Press, 1997, 89–104.

Holley, Rose, How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs, in: D-Lib Magazine 15 (2009), https://doi.org/10.1045/march2009-holley.

Horne, John/Kramer, Alan: German Atrocities, 1914: A History of Denial, New Haven: Yale University Press, 2002.

Jannidis, Fotis/Kohle, Hubertus/Rehbein, Malte (eds.), Digital Humanities: Eine Einführung, Stuttgart: Springer, 2017

Kauffmann, Kai, Wissensvermittlung, Kulturpolitik und Kriegspropaganda: Thesen zur Kriegspublizistik der deutschen Rundschauzeitschriften 1914–1918, in: Olivier Agard/Barbara Beßlich (eds.), Krieg für die Kultur? Une guerre pour la civilisation? Intellektuelle Legitimationsversuche des Ersten Weltkriegs in Deutschland und Frankreich (1914–1918), Berlin: Peter Lang, 2018, 113–128.

Latham, Sean, Scholes, Robert, The Rise of Periodical Studies, in: *PMLA* 121 (2006), 517–531.

Nicholson, Bob, Counting Culture; or, How to Read Victorian Newspapers from a Distance, in: Journal of Victorian Culture 17 (2012), 238–246.

Öhrig, Bruno, Adolf Dirr (1867–1930): Ein Kaukasusforscher am Münchner Völkerkundemuseum, in: Münchner Beiträge zur Völkerkunde 6 (2000), 199–234.

Reventlow, Ernst, Die Frage der türkischen Meerengen und ihre Entwicklung, in: Süddeutsche Monatshefte 14 (1917): 432–66.

Riddell, Allen Beye, How to Read 22,198 Journal Articles: Studying the History of German Studies with Topic Models, in: Lynne Tatlock/Matt Erlin (eds.), Distant Readings: Topologies of German Culture in the long nineteenth Century, New York: Rochester, 2014, 91–113.

Rieder, Bernhard/Röhle, Theo, Digital Methods: Five Challenges, in: David M. Berry (ed.), Understanding Digital Humanities, New York: Palgrave Macmillan, 2012, 67–84.

Roff, Sandra, From the Field: A Case Study in Using Historical Periodical Databases to Revise Previous Research, in: American Periodicals 18 (2008), 96–100.

Schaepdrijver, Sophie De, Belgium, in: John Horne (ed.), A Companion to World War I., Chichester: Wiley-Blackwell, 2010, 386–402.

Schmidt, Birgit, 'Die Frauenpflichtlerin' – Zur Erinnerung an Nadja Strasser, in: Aschkenas 16 (2006), 229–259.

Selig, Wolfram, Paul Nikolaus Cossmann und die Süddeutschen Monatshefte von 1914–1918: Ein Beitrag zur Geschichte der nationalen Publizistik im Ersten Weltkrieg, Osnabrück: A. Fromm, 1967.

Strange, Carolyn et al., Mining for the Meanings of a Murder: The Impact of OCR Quality on the Use of Digitized Historical Newspapers, in: Digital Humanities Quarterly 8 (2014), http://www.digitalhumanities.org/dhq/vol/8/1/000168/000168.html.

Straßer, Nadja, Die russische Frau in der Revolution, in: Süddeutsche Monatshefte 12 (1915), 647–652.

Tschechow, Anton, Die Gefängnisinsel Sachalin, in: Süddeutsche Monatshefte 12 (1915), 701–710.

Zechlin, Egmont, Die deutsche Politik und die Juden im Ersten Weltkrieg, Göttingen: Vandenhoeck u. Ruprecht, 1969.

Ziegesar, Jozef Haller van, Der mittlere Unterricht in Belgien, in: Süddeutsche Monatshefte 13 (1916), 605–616.