Beyond Budweiser

Creating a Digital Archive of Popular German-American Newspaper Literature

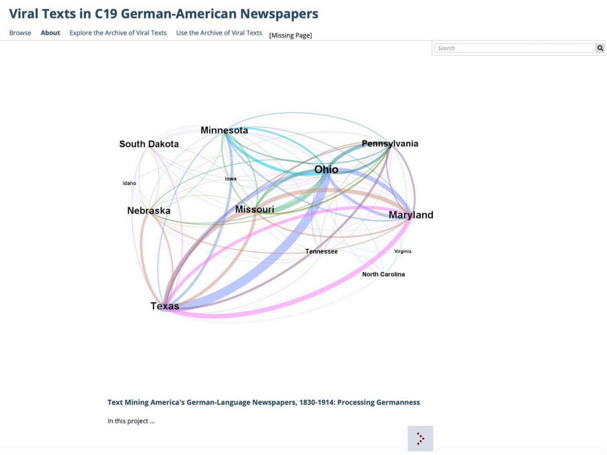
Keck, Jana
keck@ghi-dc.org
German Historical Institute Washington DC

Blessing, Andre
andre.blessing@ims.uni-stuttgart.de
Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung

Who gets to be remembered and historicized by ways of – digital – record creation? For many German-Americans in the long nineteenth century, German-language newspapers were the primary source of both information and entertainment. So far, research on the German-American press has predominantly focused on the male editors, writers, or advertisers and their influence – and success stories – on U.S. politics and economy. These HIStories have entered into schoolbooks and popular culture in the U.S. and Germany alike. Idealized versions, for instance, of the nineteenth-century German-American migrant as a hard-working, bright, self-made man have been "reused" for marketing strategies as the 2017 Budweiser Super Bowl commercial illustrates. The commercial video ad titled "Born the Hard Way" tells the fictitious tale of Adolphus Busch, co-founder of the brewery dynasty Anheuser-Busch, who emigrated to the U.S. in 1857, and ends with the slogan: "when nothing stops your dream."[1]

Such representations offer limited access to histories about the everyday life of other historical actors that go beyond the elite. Digitized historic German-language newspapers in the *Chronicling America* database[2] seem to provide a fruitful platform to find unknown stories that shed more light onto the daily experiences of historical actors of migration such as, for instance, women, girls, mothers, or daughters. Even though, such digitization projects offer access to thousands of newspapers pages (cf. Soni et al. 2021), there are no innovate methods to search and analyze them (cf. Hausdewell et al. 2020). Which keywords to enter when one does not even know what they are looking for?

In order to systematically rewrite histories about representations of marginalized groups in the German-American press, we are creating an expanded version of *Chronicling America's* repository: using OMEKA S,[3] a free, flexible, and open-source web-publishing platform, users will have the opportunity to access news content that was not only published once, but reprinted several times across states and decades (see fig 1). The dataset of reprinted texts was created with text reuse detection software (Smith et al. 2013).[4]



Text Mining America's German-Language Newspapers, 1830-1914: Processing Germanness

In this project ...

Fig. 1: Sample homepage (OMEKA S) of the dataset of reprinted texts in nineteenth-century German-American newspapers (in progress).

With this method, we have uncovered approximately 500,000 viral texts, in Ryan Cordell's words, who uses the social media metaphor to describe reprinting practices in the industrial age.[5] These viral texts were not only hard news. They range from advertisements, or factual texts to poems. However, these texts are not yet categorized into different genres. To add another layer that will make an expanded search possible by adding genre as metadata, we are using unsupervised methods (*topic models* , *clustering*) and supervised classification methods enabled by manual annotations that were integrated into a web-based interface, an adaptation of the DFR-Browser,[6] and extended by an annotation module.

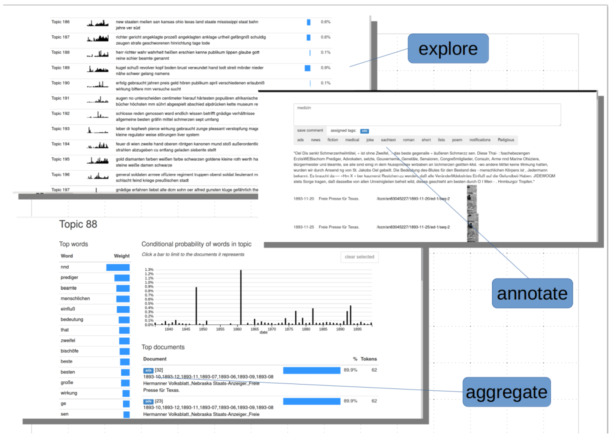


Fig. 2: DFR-Browser extended by an annotation module to tag genres

To annotate data, which is then being used as training data for the genre classifier, we have added a sample selection of documents. Such a step allows us to annotate texts, which have a high and a low likelihood of representing a specific topic.[7] As figure 2 illustrates, computationally classifying genres does not mean that a text will be 100 % categorized as belonging to one type, but in varying degrees to several types. The genre which achieves the highest score will be decisive for the dataset published in OMEKA S. With this approach, we have identified 10 different newspaper genres. By linking the annotation interface with the OMEKA site, users will be able to gain insight into genre classification process and examine similarities and anomalies between texts and genres using a mixed-methods approach (Sá Pereira 2019). Additionally, scholars can always get redirected to *Chronicling America* and to examine, for instance, where the text was embedded in the newspaper page.

In nineteenth-century newspaper ads, women were not only used as marketing strategies for medical products to cure female weakness, but predominantly for products marketed to both sexes (Keck 2021). By linking different datasets and interfaces, our project shows how we can efficiently use data as “a check-in, (...), a resource to begin and continue dialogue”[8] in gender studies. Only accessible datasets can be passed on to future generations. Adding the category of genre as metadata, provides a distinct approach to simply using keyword search, which requires prior – often biased – knowledge of the user. As Temi Odumosu proposes, we should see data and metadata as ways to rethink cataloguing spaces with the potential to alter historical imbalances of power (2020: 299). Machines can help in this way because they approach data differently: the algorithms used for text reuse detection and text classification do neither privilege specific writers, topics or groups.

Fußnoten

1. https://www.youtube.com/watch?v=IZaQQvfIfPQ
2. For details, see Chronicling America‘s “About” page.
3. https://omeka.org/s/
4. See https://github.com/dasmiq/passim.git.
5. https://viraltexts.org
6. https://agoldst.github.io/dfr-browser/
7. For a critical reflection on the use of topic modeling in the humanities, see Shadrova (2021).
8. https://www.manifestno.com

Bibliography

Hauswedell, Tessa / Nyhan, Julianne / Beals, Melodee / Terras, Melissa / Bell, Emily (2020): "Of global reach yet of situated contexts: an examination of the implicit and explicit selection criteria that shape digital archives of historical newspapers". In: *Archival Science* 20, 139-165. DOI: https://doi.org/10.1007/s10502-020-09332-1.

Keck, Jana (2021): "Let’s talk data, bias, and menstrual cramps: Voicing Gerwomanness in the ninetheenth century and today". In: *Bulletin of the German Historical Institute (Spring 2021)*. https://www.ghi-dc.org/fileadmin/publications/Bulletin/bu68/bu68_61.pdf

Odumosu, Temi (2020): "The Crying Child: On Colonial Archives, Digitization, and Ethics of Care in the Cultural Commons". In: *Current Anthropology*, vol. 61. DOI: 10.1086/710062.

Sá Pereira, Moacir P. de (2019): "Mixed Methodological Digital Humanities". In: *Debates in the Digital Humanities*, chapter 34. DOI: https://doi.org/10.5749/9781452963785.

Shadrova, Anna (2021). "Topic models do not model topics: epistemological remarks and steps towards best practices". In: *Journal of Data Mining and Digital Humanities*, Episciences.org, 2021. DOI: https://doi.org/10.46298/jdmdh.7595.

Smith, David A. / Cordell, Ryan / Maddock Dillon, Elisabeth (2013). "Infectious Texts: Modelling Text Reuse in Nineteenth-Century Newspapers". In: *Proceedings of the Workshop on Big Humanities*, 86–94. Washingyon, DC: IEEE Computer Society Press.

Soni, Sandeep / Klein, Lauren F. / Eisenstein, Jacob (2021): "Abolitionist Networks: Modeling Language Change in NineteenthCentury Activist Newspapers". In: *Journal of Cultural Analytics*, 43 (1). Doi: 10.22148/001c.18841.