Emily Moss
INFO 220
Fall 2017

Discussion Post

Week 2:

Prompt: OCR technology continues to improve our ability to machine read new fonts, alphabets, and scripted languages. However, it is not perfect. What do you think is an acceptable error rate? How can we account for human correction on large-scale digitization projects? How can we work around errors?

My response: The development of optical character recognition (shorthanded as OCR) technology to make physical texts available as digital texts made possible the rise of e-books in addition to opening up a wealth of possibilities for digital humanities research and methods. However, there remains considerable concern about the accuracy of OCR'd texts as well as debate on the significance of accuracy when working with OCR'd texts.

As Holley (2009) effectively lays out in her article on the Australian Newspaper Digitization Program, the number one factor which influences OCR accuracy is the quality of the original source as most OCR software is only able to claim 99% accuracy on new, clean images of good quality (para. 14). Scanning resolution and size of output as well as software algorithms, pattern images in software database, and in-built dictionaries all have influence on the accuracy of the OCR'd text as well. Significantly though in Holley's case study, improving upon these latter factors didn't yield much improvement on accuracy in the OCR output. The most effective means of improving accuracy in the digital texts included manual correction by community volunteers. Additionally, Holley explains how involving volunteer labor yielded benefits beyond improving the accuracy of the digital newspaper texts and minimizing cost as it also created stakeholders inside the project and furthered a sense of participatory culture within the organization.

Similarly, Strange, McNamara, Wodak, and Wood's project of text mining newspapers to better understand public sentiment around a 19th century murder trial contended with similar issues of OCR accuracy as well as "noise" in their texts (spelling variations and language variants depending on the original source). I found this phenomenon especially interesting in text-mining practices because it's one that persists today particularly in text-driven spaces like Twitter. While Holley found that volunteer manual correction was most effective, Strange et al. felt their project most benefited from the inclusion of metadata tags. In fact, the authors found that "OCR is effective in digitizing historical newspapers to roughly 80% accuracy" (para. 50) but to achieve something close to 98% accuracy requires time and labor arguably equivalent to that required to manually input the texts with corrections made during the process. Further for text mining purposes, "the cleaning was thus desirable but not essential. The addition of genre metadata led

to results of greater interest, since it allowed a focus on articles more clearly relevant to the research question" (para. 51).

This final insight gets to the debate around the importance of accuracy in OCR'd texts as when one poses the question, the short answer might be "well, it depends". For example, Kichuk (2015) argues "it is vitally important for the preservation of our print cultural heritage to ensure the quality of the digital objects hosted in digital repositories and to urge producers to build a better e-book" (p. 61). Kichuk, however, is primarily discussing e-books available through libraries or websites that people read for the content of the book itself as opposed to texts under consideration for scholarly research. In this case, it does seem that fidelity to the source material is necessary such that the reader can understand the book's content toward constructing meaning. The importance of OCR quality is also prioritized in digital humanities discussions both generally (Shahnazari, 2017) and around Natural Language Processing specifically (Piotrowski, 2017). However, there seems to be less concern with "bad OCR" in studies of stylometry toward author attribution (Wilms, 2017). Finally, as Willett (2004) states "while imperfections impede precise and complete results from keyword searches, proponents believe that with adequate OCR accuracy, the results will still prove useful, and at a significantly lower overall cost to produce than highly accurate transcriptions" (para. 20). People (scholars and otherwise) make use of different texts for different reasons and I think it's hard to divorce the purpose of engagement entirely from the significance of accuracy.

References

Holley, R. (2009, March/April). How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. *D-Lib Magazine*, *15*(3/4). Retrieved from Links to an external site.http://www.dlib.org/dlib/march09/holley/03holley.html

Kichuk, D. (2015). Loose, falling characters and sentences: The persistence of the OCR problem in digital repository e-books. *Libraries and the Academy, 15*(1), 59-91.

Piotrowski, Michael. [true_mxp]. (2017, September 12). [Tweet]. Retrieved from https://twitter.com/true_mxp/status/907598232636846080 (Links to an external site.)Links to an external site.

Shahnazari, Masoud. [CyberLiterature]. (2017, September 12). [Tweet]. Retrieved from https://twitter.com/CyberLiterature/status/907597663352344581https://twitter.com/CyberLiterature/status/907597663352344581https://twitter.com/CyberLiterature/status/907597663352344581 https://twitter.com/CyberLiterature/status/907597663352344581 (Links to an external site.)Links to an external site.

Strange, C., McNamara, D., Wodak, J., and Wood, I. (2014). Mining for the meanings of a murder: The impact of OCR quality on the use of digitized historical newspapers. *Digital Humanities Quarterly*, *8*(1). Retrieved from http://www.digitalhumanities.org/dhq/vol/8/1/000168/000168.html (Links to an external site.)Links to an external site.

Willett, P. (2004). Electronic texts: Audiences and purposes. In S. Schreibman, R. Siemen, J. Unsworth (Eds.), *A Companion to Digital Humanities.* Retrieved from http://digitalhumanities.org:3030/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-3-6&toc.depth=1&toc.id=ss1-3-6&brand=9781405103213_brand (Links to an external site.)Links to an external site.

Wilms, Lotte. [Lottewilms]. (2017, September 11). [Tweet]. Retrieved from https://twitter.com/Lottewilms/status/907232181726203904 (Links to an external site.)Links to an external site.