Women’s Writing in the Eighteenth Century: Evaluating ‘Representative’ Corpora

Original title: Canonical Corpora: What Makes a ‘Reliable’ or ‘Representative’ Source?

When we study literature through text mining, our conclusions are not based on an examination of “literature” itself, but on the corpus which is a sample or model of the imagined whole of “literature.” Despite the importance of corpus-building to the interpretation of text mining research, it is often extremely difficult to know what is in a corpus. Even large institutional resources used by many scholars provide little context for their choices of what to include or exclude. These hidden choices are particularly problematic when historical selection factors might have led to the creation of corpora which re-create social inequalities.

As an eighteenth century scholar, I examine five corpora which are used as the basis of most eighteenth century distant reading: the English Short Title Catalogue database; Eighteenth Century Collections Online; the Eighteenth Century Collections Online Text Creation Partnership; Project Gutenberg; and HathiTrust. I manually evaluate each corpus’s holdings for a narrow sample of texts, works published in England 1789-99, to investigate and contextualize their representation of female authors. For this eleven-year period, the English Short Title Catalogue provides basic bibliographic data for nearly 52,000 titles, but the Eighteenth Century Collections Online Text Creation Partnership corpus of XML-encoded full texts includes fewer than 500 titles. This difference raises the question: why were the other 51,500 titles not considered worth the investment of scholarly effort? And with particular urgency: do the most invested-in resources underrepresent women?

In addition to calculating the gender ratio of authorship for each of the five corpora (including an assessment of the unsigned or unattributable works), I correlate gender with the basic categories of writing present in the corpus, to form more specific conclusions about the role of gender in corpus-building. Simple topic-modelling of the titles of eighteenth century works allows me to identify broad categories such as plays, poetry, pamphlets, novels, songs, sermons, ephemera, and so on. Titles in the eighteenth century are long, and served as advertisements, intentionally created to actively communicate this level of information about their works’ contents: although topic modelling of titles would not be sufficient in other centuries, it is remarkably effective here. Identifying the representation within each corpus of these categories of writing reveals a predictable preference for “literary” forms such as novels and poetry in the smaller corpora.

This preference for particular kinds of writing might explain changes in gender representation of smaller corpora. If the novel is the domain of women, for example, a corpus can underrepresent women by underrepresenting novels. Or it could include a representative number of novels, but disproportionately include novels by men. My investigation allows me to identify the patterns of exclusion. Asking bibliographical questions of multiple corpora, in order to learn about the corpora themselves, emphasizes an under-examined stage of text mining research, and provides a basis for other scholars to use these corpora more precisely. Asking pointed questions about gender, in particular, may allow future users of these corpora to address some impacts of historical sexism.