Reconstructing Readerly Attention: Citational Practices and the Canon, 1789–2016

Mark Algee-Hewitt
mark.algee-hewitt@stanford.edu
Stanford University, United States of America

David McClure
dclure@stanford.edu
Stanford University, United States of America

Hannah Walser
walser@stanford.edu
Stanford University, United States of America

In its ability to extract feature sets, relate texts within an abstract space, and semantically parse groups of texts, computational textual analysis has functioned primarily as a formalist intervention into literary study. When practitioners venture outside of the formal features of the texts themselves, it is author or date that serves as the point of contact between the text and its wider context. And yet, texts offer a rich history of reception: as different interpretive communities (Fish 1980) receive and reinterpret novels, poems or plays, they recontextualize the literary object to suit the particular socio-cultural goals of their period or nationality. Lacking detailed accounts of reading practices at large scales, even traditional practitioners of literary history have been unable to reconstruct the history of reception of even the most historically canonical texts. In this project, we leverage the ability of Digital Humanities to recover, at least provisionally, a large-scale history of textual reception by exploring the patterns of citation that reveal the attention paid to specific texts across their history as readerly objects. How are certain canonical texts cited over time and what can the attention paid to different segments of texts with a rich reception history tell us about the reading or social practices of different historical periods? How do different groups of readers (particularly authors and critics) quote text differently as they make use of passages in their own writing? And how do specialists and non-specialists cite the same text differently? By exploring the locus of attention within a canonical text, both across groups of readers and across history, we provisionally reconstruct a historically and socially contingent map of a text's reception history.

As Piper and Algee-Hewitt have argued in "The Werther Effect" (Piper and Algee-Hewitt 2014), practices of citation, the embedding of the language of a text within other works, can reveal patterns of reception even within a single author's corpus. In this project, we expand this approach multi-dimensionally, identifying passages of canonical works quoted in other novels, in critical articles by specialists and non-field specialists, and in a larger undifferentiated corpus of text. The scale of our analysis enables us to identify what parts of a text have received the most writerly attention overall and how that attention has been shaped over time. By moving from semantics to passages, we switch our attention from the intangible metrics of semantic similarity, to specific, quotation-level instances of citation that demonstrate specific attention to identifiable parts of our target texts. Drawing on work in sequence alignment by David Smith et al (2013) and Richard So et al. (forthcoming), we will therefore be able to explore patterns of attention that have been paid to a text by identifiable groups of readers. We argue that these patterns of citationality serve as a proxy for the reception of a text: while necessarily limited to readers who themselves were authors (or critics), they nevertheless represent an important category of reception available to analysis.

To extract the quotations, we used Python's "difflib" module, wrapped up as a parallelized MPI program that runs on an HPC cluster. To compare any two individual texts - for example, when checking for passages from Hamlet inside of a novel from the Gale American Fiction corpus - the texts are first split into tokens and passed through a filter that removes a set of 200 stopwords. This speeds up the alignment algorithm (the high-frequency words that get pulled out make up a significant portion of the total words in any given text, producing shorter sequences) and also has the advantage of making the alignment process less sensitive to small changes in function words, which seem to get shuffled around or changed fairly frequently when a text is quoted. For example, a change from:

> And crook the pregnant hinges of the knee
> **Where** thrift may follow fawning

to

And crook the pregnant hinges of the knee
That thrift may follow fawning

still gets picked up as a quotation, since the semantically significant words - crook, pregnant, hinges, knee, thrift, fawning - stay the same.

These filtered sequences of tokens then get passed through the alignment algorithm, which produces a set of matches, recoded in terms of their starting positions in each text and the length of the matching subsequence. To ensure that the matches represent actual quotations, we discarded matches shorter than 5 tokens (not counting stopwords), since alignments shorter than this include a fair number of false positives, generic word sequences that likely don't represent any kind of meaningful quotation or intertextuality - for example, many are numbers, things like "five hundred thousand." This gives us high "precision" - almost all of the alignments that are included in the final analysis represent legitimate quotations to the play - but it also drops down the "recall" somewhat, since some of the shorter alignments are, in fact, real quotes - things like "weighing delight and dole." We are currently evaluating a couple of strategies for identifying these alignments that are short but semantically "focused" enough that we can say with confidence that they should be included in the set of valid matches.

We use this method of sequence alignment to trace the quotations of five canonical texts with a rich citation history across four corpora. Our selected texts include Shakespeare's *Hamlet*, Milton's *Paradise Lost*, Dickens' *A Christmas Carol*, Caroll's *Alice in Wonderland* and Wordsworth's *Prelude*. Not only are all of these texts heavily quoted by critics, but, we argue, they have entered the literary and cultural consciousness of both Britain and America such that the passages that are cited by authors and critics reveal interpretive and readerly practices both across time and between different groups. For each text, we extract all of the citations from it that are five words (excluding stopwords) or longer, that occur in each of our four corpora: the full-text Hathi trust corpus, representing a massive sample of writing in the nineteenth and early twentieth centuries; a literature-specific corpus of 28,000 novels from England and America dating from 1789-2016; and two corpora of articles on literary studies, one a corpus of 10 journals of literary criticism and history (e.g. *PMLA*, *NLH*, *Critical Inquiry*) and one a corpus of 10 field-specific journals focused specifically on the authors represented in our group of canonical texts (e.g. *Shakespeare Quarterly*, *Milton Studies*, *Wordsworth Circle*).

Between these four corpora, we are able to differentiate the kind of attention paid to our primary canonical texts by four different groups of readers. Do novelists pay attention to different parts of a text than authors in general in the nineteenth and twentieth centuries? Do general literary critics quote different parts of *Hamlet* than Shakespeare specialists? And for each of these corpora, how does the citation map of each of our texts change over time?

For example, a citation map of *Hamlet* in our novel corpus revealed 1,693 quotations of five or more non-stopword tokens, which collectively cover about 25% of all words in the play. When we plot the frequency of citations across the narrative of the drama (broken into 500 bins), our method reveals the passages most quoted by novelists of the nineteenth and twentieth centuries (Figure 1). From this citation map, we can see that Hamlet's soliloquy ("to be or not to be") is among the top three passages cited by novelists; however, the quotation that clearly dominates the use of *Hamlet* by this group of readers comes from Act 5, Scene 2 "There's a divinity that shapes our ends, / Rough-hew them how we will.—"

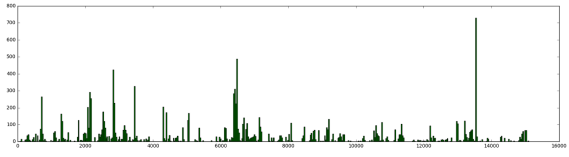

Figure 1 Numbers of citations of Hamlet in nineteenth and twentieth-century novels. Each passage is 1/500 of the text.

Although not among the most identifiable passages today, this quotation clearly had a resonance for the readers of the nineteenth and early twentieth centuries.

By comparing this map of citations across the narrative of Shakespeare's plays to ones that are both drawn from our comparative corpora and periodized across the two centuries they represent, we are able to show how different passages gain and lose meaning across time and between kinds of reading. As we expand this to all five of our canonical texts read into all four of our corpora, we can shed light on how historically and disciplinarily specific practices of reading shaped the horizons of interpretation for specific works, and begin to reconstruct these reader-based practices in ways that are open and tractable to the Digital Humanities.

Bibliography

Fish, S. (1980) *Is there a text in this class?* Cambridge: Harvard UP.

Piper, A., and Algee-Hewitt, M. (2014). "The Werther Effect I: Goethe, objecthood and the handling of knowledge." *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*. Ed. Matt Erlin and Lynn Tatlock. Rochester: Camden House. 155-184.

Smith, D., Cordell, R., Dillon, E. M. (2013). "Infectious texts: modeling text reuse in nineteenth-century newpapers." *Proceedings of the IEEE International Conference on Big Data*. 86-94.

So, R. J.; Long, H, Yuancheng, Z. (forthcoming). "The Dark Code: Modeling White-Black Literary Relations, 1880-2000. *Forthcoming.*