Learning to Read Data: Bringing out the Humanistic in the Digital Humanities

Ryan Heuser, Long Le-Khac

Victorian Studies, Volume 54, Number 1, Autumn 2011, pp. 79-86 (Article)

Published by Indiana University Press

Learning to Read Data: Bringing out the Humanistic in the Digital Humanities

Ryan Heuser and Long Le-Khac

As humanists, we believe we are trained as expert readers, able to read almost any kind of text closely, deeply, and critically. But on 21 February 2010, the two of us sat staring at a computer screen dumbfounded by a kind of text that for once we had no idea how to read. Working with a corpus of thousands of digitized nineteenth-century British novels, we had just produced a plot, similar to figure 1, showing the usage trends of a massive group of abstract words relating to social values. The plot seemed to show all these words disappearing over the nineteenth century. What exactly were we seeing here? A heated discussion of principles in the fallout of the French Revolution? A death of values in the Victorian period?

The moment was emblematic of how it can feel to encounter digital humanities work. In facing a radically new kind of text, a different kind of evidence, tremendous excitement and real anxiety mix. It's easy to understand the excitement. These emerging methods promise ways to pursue big questions we've always wanted to ask with evidence not from a selection of texts, but from something approaching the entire literary or cultural record. Moreover, the answers produced could have the authoritative backing of empirical data. But it's also understandable how these possibilities could be unsettling. By offering an entirely different model of humanities scholarship, the digital humanities raise many questions. What do we do with this kind of evidence? Can we leverage quantitative methods in ways that respect the nuance and complexity we value in the humanities? Behind these questions is perhaps a deeper concern. Under the flag of interdisciplinarity, are the digital humanities no more than the colonization of the humanities by the sciences?

While doing digital humanities research over the past two years, we have constantly wrestled with these questions. We have learned,

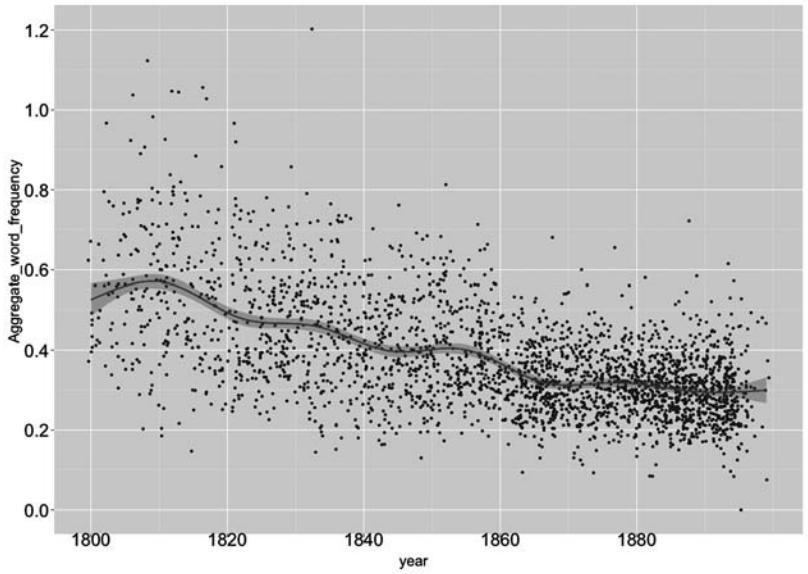

Fig. 1. Aggregate word frequency of abstract values field in percentage of total word usage, 1800–1900. Note: In these plots, each point represents the word frequency in one novel in our corpus. A local polynomial regression fit is superimposed.

happily, that the answers may not be as troubling as they can seem. We are convinced that, when done well, such research can deliver scale, empirical rigor, and the nuance the humanities value. Yet this will require deep reflection as these methods develop. More importantly, this work will depend on humanistic methods. We hope to substantiate these claims by addressing some key methodological questions and presenting a case study from our own research.

Methodological Anxieties

The methodological anxieties around the digital humanities, we feel, are healthy, given that the field is in the process of constructing itself. Indeed, when we look closely at how current digital humanities work pursues the promises of scale and empiricism, there remain many problems to work out if we are to deliver on those promises. We'll focus here on three pervasive problems: anecdotal evidence, validation, and interpretation.

Moving from anecdotal to large-scale evidence is not as straightforward as it seems. Even with millions of texts, the evidence generated

can still be anecdotal. This problem can be understood using two terms we have found useful in thinking about evidence in the digital humanities: signal and concept. A signal is the data from the feature actually being measured computationally. A concept, however, is the phenomenon we take a signal to stand for. In the digital humanities, the interest and impact of our arguments are based on concepts, but computers can only measure signals, which are always smaller than concepts. For example, we may want to explore the waning of cultural memory, but have as our data only trends in the mentions of particular years (for instance, trends in how many times "1930" or "1945" is used) (Michel et al. 178). The essential problem of quantitative evidence, then, is in deciding how to bridge the perpetual distance between the signals we have and the concepts we want them to represent. To reduce this distance, we must work not only on the size of our corpus but on the robustness of the signals we cull from it.

Because it's difficult to bridge this distance, a second problem emerges: we tend to read data in terms of the concepts we already have at hand. This can be seen at work when Dan Cohen, in "Searching for the Victorians," presents frequency plots of several words related to faith and concludes that his data confirms the Victorian crisis of faith (and thus seems reliable). Even a visual inspection of the plots, though, reveals that these words do not share any single, declining trend. Many of the words presented have distinct and complex trends, some rising in the Victorian era before declining. In such cases, a familiar concept is applied too hastily to the data, thus flattening the data's nuances and complexities. A troubling corollary to this is a tendency to throw away data that does not fit our established concepts. When Cohen discards a striking correlation between "belief," "atheism," and "Aristotle" as an accident of the data, he does just this. Whether or not the correlation is accidental should be decided by statistical analysis rather than the feeling that it doesn't make sense. If we required all data to make sense—that is, fit our established concepts—quantitative methods would never produce new knowledge. If the digital humanities are to be more than simply an efficient tool for confirming what we already know, then we need to check this tendency to seek validation.

Indeed, making the leap from signal to concept is tricky. But, as we've begun to suggest, there are ways to build signals that allow us to make this leap with greater confidence. To this end, we see two key goals when building signals: maximizing scale and maximizing conceptual

coherence. Scaling up the number of word frequencies being tracked, for instance, helps limit the number of possible interpretations. Imagine finding not several but hundreds of words sharing one trend. As the number of words undergoing the same change increases, the number of causes that could plausibly have affected *all* of those distinct words decreases. Maximizing a signal's conceptual coherence helps in making potential causes identifiable. While acknowledging the possibility of random coincidence, we can still assume that if causes acted collectively on all of the words sharing a trend, it's probably because the words share some common links that are relevant in explaining why they were acted upon. Thus, the clearer the links between words, the easier it is to identify the probable cause by examining those links.

Case Study: The Languages of Social Space in the Nineteenth-Century British Novel

Let's look now at some specific strategies for addressing these methodological questions.[1] We present here parts of our own research analyzing language developments in 2,779 nineteenth-century British novels as a case study of quantitative methods in action. We focus on word frequency trends much as have Dan Cohen and others. Given the richness of language and the diffuseness of cultural trends, it's unlikely that the latter would be isolatable to, or precisely measurable in, the behavior of individual words or small word groups. For this reason, and in order to capitalize on the advantages of scale and conceptual coherence, we aggregated words into semantic fields, large groups of semantically related words. In doing so, we moved beyond hand-picking words toward an unsupervised approach where computation reveals fields in the data. To achieve scale and coherence, we developed a two-pronged method of building semantic fields. We built a tool we call Correlator to find large groups of words that share a common historical trajectory. To ensure semantic coherence and interpretability, we turned to the *OED*'s historical thesaurus to identify the semantic content of the large word cohorts found through Correlator, to subdivide them into specific semantic fields, and to continue expanding them with more words.

This method revealed two massive semantic fields with striking trends. The first field, representative of what we call "abstract values," includes over 300 words, such as "moderation" and "excess," "virtue"

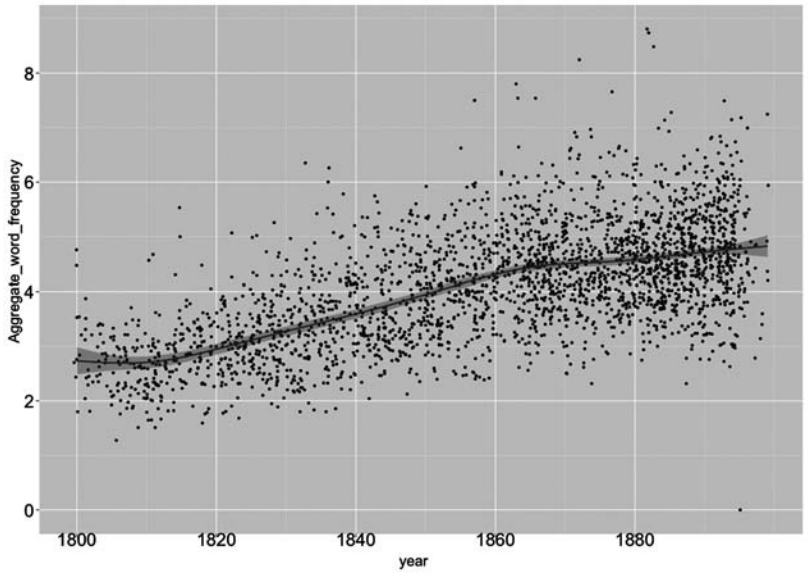

Fig. 2. Aggregate word frequency of "hard seed" field in percentage of total word usage, 1800–1900.

and "sin," "passion" and "sensibility," "disinterested" and "prejudiced." Highly polarized abstractions, these words typically relate to social norms and the regulation of behavior. The usage of this field declines precipitously over the nineteenth century (see fig. 1). The second field, which we named "hard seed" after its seed word, was even bigger, consisting of over 600 words of a variety of types: colors, body parts, numbers, locational prepositions ("down," "out," "over"), action verbs, and physical adjectives ("hard," "rough," "flat"). In contrast to the abstract values words, these are concrete, physical, specific, and non-evaluative. And again in contrast, the usage of hard seed words rises dramatically (see fig. 2).

From Numbers to Meaning

What might all this mean for the history of the nineteenth-century British novel? More importantly, how can we extract meaning from the data?

Noticing a strongly inverse relationship between the two fields, we ranked the novels in our corpus by their concentration of the two fields, producing a spectrum. This allowed us to see the data through

units familiar to us as literary scholars, the actual novels, genres, and authors in our corpus. Instead of trying to interpret the frequency behaviors of semantic fields, a rather abstract object, the spectrum allowed us to understand the shifts in novelistic language more directly as changes in the kinds of novels being written. We found this to be a key tactic in moving from signals to concepts: translating and visualizing data into readily interpretable forms.

From left to right, the spectrum ranked novels with the highest frequency of abstract values words to the lowest, and conversely, the lowest frequency of hard seed words to the highest. This produced a distribution that began with the evangelical novel (highest in abstract values, lowest in hard seed), closely followed by the Gothic, then by Austen, Scott, and Eliot. Toward the right were the urban and industrial novel and Dickens. At the extreme right (lowest in abstract values, highest in hard seed) were adventure novels, fantasy, science fiction, and children's literature.

What would this intriguing distribution reveal if we could read it? To tackle this problem of interpretation required a powerful, but unintuitive, strategy: to move from numbers to meaning, what may be needed is more numbers. As humanists, we may fear that gathering more quantitative data only moves us farther from the qualitative meaning we seek, but we're suggesting that having more kinds of data actually moves us closer to finding meaning. When one has a single dimension of data, it's too easy to leap in almost any direction and to any number of interpretations. For example, when at first we only had data showing the decline of the abstract values fields, we guessed that we were seeing a shift from late eighteenth-century values to Victorian values. When we had to square our hunch with additional data, the inversely related trend of hard seed words, a very different picture emerged. With the addition of the spectrum, we had several rich and suggestive datasets to triangulate.

We began with the characteristics of the abstract values—highly polarized, evaluative words related to norms of social regulation—and mapped them onto the spectrum, which allowed us to see that the movement marks a change in the social space of the novel. Social spaces expand from small and constrained to wider and freer, from the tight communities of the evangelical or village novel to the city or beyond, to exotic islands and fantasy worlds. We now had three corresponding dimensions of data: a decline in abstract values words, a rise in hard seed words, and, as these temporal trends unfold, an expansion of social space in the novel.

Things began to fall into place. We could see now the fit of the abstract values fields to small, constrained social spaces, which can be thought of in Raymond Williams's terms as "knowable communities" (165). A unified and clear-cut system of social values would undergird the legibility characteristic of a knowable community. That the abstract values fields so suited to small social spaces decline over the century led us to consider whether this could be the result of the deep changes in social space stemming from urbanization in Britain. With radical increases in the scale and complexity of forms of social organization in this period, the decline in abstract values words could reflect their obsolescence for making sense of this new, less knowable kind of society.

But apparently, the hard seed fields found a natural home there. This helped us realize that the concrete, physical hard seed words would be well suited to render the disorientation and variety of the city and more complex social spaces. With stable social schema unavailable for making sense of relations, people, and experience, the individuals inhabiting these spaces are instead faced with opaque physical detail. The panoply of strangers in the crowded city street cannot be known individually, but only as a succession of anonymous bodies and surfaces. Concrete language with its enormous range and nuance can render this proliferating and variegated social system.

In the end, we learned that the open-ended process of interpretation is made more rigorous by having to account for a wide set of related observations. This approach can be summarized as a hypothesis-testing mode of interpretation. Engaged in a constant dialogue between evidence and interpretation, hypothesis testing seeks to eliminate potential theories by testing them against multiple forms of data, resulting in a stronger argument. We believe this model, when combined with robust, carefully culled data, can provide a new and powerful form of investigation for humanities scholarship.

In concluding this case study, though, it's crucial to emphasize that from semantic taxonomies to knowledge of individual texts and authors, to contextualization in nineteenth-century cultural history, every step of this research has been deeply informed by the knowledge and methods of the humanities, without which we would have remained as dumbfounded as we were in first encountering that plot of abstract values words.

Stanford University

NOTES

This research was made possible in part by funding from a Mellon Foundation grant.

1. For a fuller discussion of our research methods and results than is possible here, please see http://litlab.stanford.edu/semanticcohort.

WORKS CITED

Cohen, Dan. "Searching for the Victorians." *Dan Cohen's Digital Humanities Blog*, 4 Oct. 2010. Web. 24 Jan. 2011.

Michel, Jean-Baptiste, et al. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* (14 Jan. 2011): 176–82.

Williams, Raymond. *The Country and the City*. New York: Oxford UP, 1973.