PART IV][Blog Posts

Link to website: https://dhdebates.gc.cuny.edu/read/untitled-88c11800-9446-469b-a3be-3fdb36bfbd1e/section/402e7e9a-359b-4b11-8386-a1b48e40425a

Text: A Massively Addressable Object

MICHAEL WITMORE

At the Working Group for Digital Inquiry at Wisconsin, we've just begun our first experiment with a new order of magnitude of texts. Jonathan Hope and I started working with thirty-six items about six years ago when we began to study Shakespeare's First Folio plays (Witmore and Hope). Last year, we expanded to three-hundred and twenty items with the help of Martin Mueller at Northwestern, exploring the field of early modern drama. Now that the University of Wisconsin has negotiated a license with the University of Michigan to begin working with the files from the Text Creation Partnership (TCP), which contains over twenty-seven thousand items from early modern print, we can up the number again. By January, we will have begun our first one-thousand item experiment, spanning items printed in Britain and North America from 1530 through 1809. Robin Valenza and I, along with our colleagues in computer sciences and the library, will begin working up the data in the spring. Stay tuned for results.

New experiments provide opportunities for thought that precede the results. What does it mean to collect, tag, and store an array of texts at this level of generality? What does it mean to be an "item" or "computational object" within this collection? What is such a collection? In this post, I want to think further about the nature of the text objects and populations of texts we are working with.

What is the distinguishing feature of the digitized text—that ideal object of analysis considered in all its hypothetical relations with other ideal objects? The question itself goes against the grain of recent materialist criticism, which focuses on the physical existence of books and practices involved in making and circulating them. Unlike someone buying an early modern book in the bookstalls around St. Paul's four hundred years ago, we encounter our TCP texts as computational objects. That doesn't mean that they are immaterial, however. Human labor has transformed them from microfilm facsimiles of real pages into diplomatic quality digital transcripts, marked up in TEI so that different formatting features can be distinguished. That labor is as real as any other.

What distinguishes this text object from others? I would argue that a text is a text because it is *massively addressable at different levels of scale*. Addressable here means that one can query a position within the text at a certain level of abstraction. In an earlier post, for example, I argued that a text might be thought of as a vector through a metatable of all possible words (Witmore). Why is it possible to think of a text in this fashion? Because a text can be queried at the level of single words and then related to other texts at the same level of abstraction: the table of all possible words could be defined as the aggregate of points of address at a given level of abstraction (the word, as in Google's new Ngram corpus). Now, we are discussing ideal objects here; addressability implies different levels of abstraction (character, word, phrase, line, etc.), which are stipulative or nominal: such levels are not material properties of texts or Pythagorean ideals; they are, rather, conventions.

Here's the twist. We have physical manifestations of ideal objects (the ideal *1 Henry VI*, for example), but these manifestations are only provisional realizations of that ideal. (I am using the word manifestation in the sense advanced in the Online Computer Library Center's Functional Requirements for Bibliographic Records [FRBR] hierarchy.[1]) The book or physical instance, then, is *one of many levels of address*. Backing out into a larger population, we might take a genre of works to be the relevant level of address. Or we could talk about individual lines of print, all the nouns in every line, every third character in every third line. All this variation implies massive flexibility in levels of address. And more provocatively, when we create a digitized population of texts, our modes of address become more and more abstract: all concrete nouns in all the items in the collection, for example, or every item identified as a "History" by Heminges and Condell in the First Folio. Every level is a provisional unity: stable for the purposes of address but also stable because it is the object of address. Books are such provisional unities. So are all the proper names in the phone book.

The ontological status of the individual text is the same as that of the population of texts: both are massively addressable, and when they are stored electronically we are able to act on this flexibility in more immediate ways through iterative searches and comparisons. At first glance, this might seem like a Galilean insight, similar to his discipline-collapsing claim that the laws that apply to heavens (astronomy) are identical with the ones that apply to the sublunar realm (physics). But it is not.

Physical texts were *already* massively addressable before they were ever digitized, and this variation in address was and is registered at the level of the page, chapter, the binding of quires, and the like. When we encounter an index or marginal note in a printed text—for example, a marginal inscription linking a given passage of a text to some other in a different text—we are seeing an act of address. Indeed, the very existence of such notes and indexes implies just this flexibility of address.

What makes a text a text—its susceptibility to varying levels of address—is a feature of book culture and the flexibility of the textual imagination. We address ourselves to this level, in this work, and think about its relation to some other. "Oh, this passage in *Hamlet* points to a verse in the Geneva bible," we say. To have this thought is to dispose relevant elements in the data set in much the same way a spreadsheet aggregates a text in ways that allow for layered access. A reader is a maker of such a momentary *dispositif* or device, and reading might be described as the continual redisposition of levels of address in this manner. We need a phenomenology of these acts, one that would allow us to link quantitative work on a culture's "built environment" of words to the kinesthetic and imaginative dimensions of life at a given moment.

A physical text or manifestation is a provisional unity. There exists a potentially infinite array of such unities, some of which are already lost to us in history: what was a relevant level of address for a thirteenth-century monk reading a manuscript? Other provisional unities can be operationalized now, as we are doing in our experiment at Wisconsin, gathering one thousand texts and then counting them in different ways. Grammar, as we understand it now, affords us a level of abstraction at which texts can be stabilized: we lemmatize texts algorithmically before modernizing them, and this lemmatization implies provisional unities in the form of grammatical objects of address.

One hundred years from now, the available computational objects may be related to one another in new ways. I can only imagine what these are: every fourth word in every fourth document, assuming one could stabilize something like "word length" in any real sense. (The idea of a word is itself an artifact of manuscript culture, one that could be perpetuated in print through the affordances of moveable type.) What makes such thought experiments possible is, once again, the addressability of texts as such. Like a phone book, they aggregate elements and make these elements available in multiple ways. You could even think of such an aggregation as the substance of another aggregation, for example, "all the phone numbers belonging to people whose last name begins with A." But unlike a phonebook, the digitized text can be reconfigured almost instantly into various layers of arbitrarily defined abstraction (characters, words, lines, works, genres). The mode of storage or virtualization is precisely what allows the object to be addressed in multiple ways.

Textuality *is* massive addressability. This condition of texts is realized in various manifestations, supported by different historical practices of reading and printing. The material affordances of a given medium put constraints on such practices: the practice of "discontinuous reading" described by Peter Stallybrass, for example, develops alongside the fingerable discrete leaves of a codex. But addressability as such: *this* is a condition rather than a technology, action, or event. And its limits cannot be exhausted at a given moment. We cannot, in a Borgesian mood, query all the possible data sets that will appear in the fullness of time. And we cannot import future query types into the present. But we can and do approximate such future searches when we automate our modes of address in unsupervised multivariate statistical analysis—for example, factor analysis or Principle Component Analysis (PCA). We want all the phonebooks. And we can simulate some of them now.

NOTES

This chapter originally appeared as "Text: A Massively Addressable Object" (http://winedarksea.org/?p=926).

1. http://www.oclc.org/research/publications/library/2003/lavoie_frbr.pdf.

BIBLIOGRAPHY

Hope, Jonathan, and Michael Witmore. "The Hundredth Psalm to the Tune of 'Green Sleeves': Digital Approaches to Shakespeare's Language of Genre." *Shakespeare Quarterly* 61, no. 3 (2010): 357–90.

Witmore, Michael. "Texts as Objects II: Object Oriented Philosophy. And Criticism?" *Wine Dark Sea*. September 17, 2009. http://winedarksea.org/?p=381.

NEXT CHAPTER

Blog Post: The Ancestral Text | Michael Witmore

PREVIOUS NEXT