Distant Reading

A case of close reading

We can learn a great deal by comparing different versions of a text. For instance, the opening of the novel Mrs. Dalloway contains what is perhaps Virginia Woolf's most well-known first line, but in an **earlier short story** she wrote about Clarissa Dalloway a very similar line differed by one word:

Mrs. Dalloway said she would buy the gloves herself. (Virginia Woolf, "Mrs. Dalloway in Bond Street," 1923)

Mrs. Dalloway said she would buy the flowers herself. (Virginia Woolf, Mrs. Dalloway, 1925)

Why is it significant that Woolf originally had Mrs. Dalloway buying gloves rather than flowers?

Literary critics, **fascinated** with these kinds of version changes, have written a great deal about what the revision meant for **the opening atmosphere of the novel** and for the novel's **plot** and **character development**. With this particular change, Woolf created the opportunity to begin the action with a **journey to a florist** and to weave the **motif of flowers** throughout the novel.

It is easy to see a difference in an opening line and to consider what it reveals about an author's creative process.

Literary scholars often employ a technique called **'close reading'**. The scholar **carefully analyzes a text** sentence-by-sentence and then **draws conclusions about the text as whole**.

As another example, let's take a look at the first few sentences of ***Pride and Prejudice*** **(1813) by Jane Austen**:

"It is a truth universally acknowledged, that a single man in possession of a good fortune must be in want of a wife. However little known the feelings or views of such a man may be on his first entering a neighborhood, this truth is so well fixed in the minds of the surrounding families, that he is considered as the rightful property of some one or other of their daughters"

--> Who has read Pride and Prejudice and cares to close read this passage? How does it set the stgae and the tone for the novel?

A close reading of this passage will unveil **the premise of the entire novel**.

- Austen writes: "It is a truth universally acknowledged...." Based on just these first few words, the scholar can conclude that in the world of *Pride and Prejudice*––the world of the eighteenth-century upper-middle class––**everyone agrees on something**, and that something is most likely the main point Austen is trying to make.
- Austen goes on to explain what this truth is: "that a single man in possession of a good fortune must be in want of wife." Thus, the scholar draws the conclusion that **the book will discuss marriage and money**.
- Austen continues: "this truth is so well fixed in the minds of the surrounding families, that he is considered as the rightful property of some one or other of their daughters." The scholar would then

note that the novel discusses a **single man (Darcy)**, in possession of a **good fortune, (10,000 pounds a year)**, who ultimately marries the protagonist, Lizzie Bennet.

- The literary scholar will also note the unusual use of the word “property” in respect to a man––**since when is a man with money considered property? At this time, women were not even allowed to inherit money, and so were basically given away by their fathers to their husbands.** Austen, by conflating rich, single men with property, plays on the traditional marital roles and **sets the stage for the (surprisingly) matriarchal society she portrays**.

Thus, the traditional view is that the literary scholar can “close read” only the first lines of Pride and Prejudice and already draw conclusions about the topics the novel may address. This is the act of “close reading,” a method of literary analysis that focuses on the details of a text.

Around the turn of the 21th century, the phrase "close reading" was ofen put in comparison and opposed to "distant reading". Today, we will discuss the tension that exists between these two methods of (literary) analysis, and whether or not they are mutually exclusive.

The rise of distant reading

The term **'distant reading' resonates across DH**: It is played on in book titles (Distant Horizons, Underwood 2019) and adapted to new fields ('Distant Viewing', Arnold and Tilton 2019). It spurs alternative formulations ('Scalable Reading', Mueller 2012) and is present in mainstream media ("What is Distant Reading?", Schulz 2011). It is a **popular and integrating** term, but can take **very specific meaning as well**.

The term distant reading was invented in about **2000 by Franco Moretti**, the scholar who was central to its development for literary study.

Distant reading is the idea of processing content -- subjects, themes, persons, or places -- or information about publication date, place, author, or title in a large number of textual items without engaging in the reading of the actual text. (Definition by Drucker, deliberately put here, because it is very broad + she is a well-known scholar in the field of DH + I think she gets it wrong, distant reading is not necesarily about "big data" but about "distant" or "indirect" reading, she creates an opposition that is not there).

However, the **semantic adoption and/or adaption of the term in research is often unclear**. Recent debates have challenged some of the assumptions of 'distant reading'.

Also, the ambiguity but also the likeness (polysemy) to the term 'close reading', have contributed to **misunderstandings** in these debates.

- Debates about distant reading range from the suggestion that it is a **misnomer** to call it reading, since it is really statistical processing and/or data mining,
- to **arguments about the size of the corpus** that is appropriate for both close and distant reading, when it comes to literary or historical (or other) works in the humanities (Underwood 2017).

  When the Italian literary scholar Franco Moretti published his book, *Graphs, Maps, Trees: Abstract Models for Literary History* in 2005, this **created quite a stir among literary scholars**. It created quite a stir because Moretti called upon people like myself to abandon the practice they’re most used to, namely close reading.

> Instead of embarking on painstaking analyses of the semantic and syntactic intricacies of single literary texts, Moretti called upon people to mine huge databases that contained thousands of literary texts, to identify recurring patterns and large-scale historical developments across national borders, and over whole centuries.

The main question that we're left with here is: Could texts be “read” at a scale that exceeded human capacity? Is reading in distant reading reduced to a form of data mining that allows information in the text or about the text to be processed and analyzed.

Therefore, **our aim is to recover the historicity of the term 'distant reading'**, the actuality of how it was originally conceived, by **delineating how its meaning has changed over time and reconstructing some of the key theoretical assumptions it carries both as a term, a concept and a practice.**

Tracing it back

This will be our guiding thesis, not claiming this is right,

Franco Moretti's concept of "distant reading" is indebted to digital technologies and is transforming the act at the heart of literary criticism: Reading.

Point 1 : Distant Reading

In order to understand how reading as a concept has changed because of Distant Reading two questions must first be addressed.

1. What is the definition of Distant Reading?
2. How does Distant Reading work?

1. What is the definition of Distant Reading?

- Schulz: "understanding literature not by studying particular texts, but by aggregating and analyzing massive amounts of data."
- Moretti: "Distant Reading: where distance... is a condition of knowledge: it allows you to focus on units that are much smaller or much larger than the text: devices, themes, tropes - or genres and systems.

World literature! Moretti devised the theory of distant reading as a practice to **"look beyond the canon"**, of 200 or so commonly close read books. This is because he felt **the need to give concrete evidence to the notion of world literature**.

He also claims that close reading will not be usable if one wishes to focus on the concept of world literature since close reading was made specially to focus in on the small number of works in the commonly accepted literary canon.

In order to study world literature he claims, we must learn how not to read.

Yet, how are close reading and distant reading different?

Close and Distant Reading

Close Reading is the act of analyzing a small set of works based upon content and structure. For example, one could analyze a Dickinson poem based on the locations of line breaks, the meanings of each word, and comparisons drawn with similar works. Another example would be to compare and contrast two characters from two different works based upon the content of their interactions with others.

Comparatively, distant reading focuses less on interpreting a small body of works in this way and instead is used to look at such things as how many books were written per year over a set number of years or who various characters in a book interact with minus the content of those interactions.

2. How does distant reading work?

Taking al this into account the current definition of distant reading is that it is used to **analyze more than the content and structure of a few books**.

Yet, this doesn't fully cover **how** this is done, **why** this is done, or **how this shows the evolution of the concept of reading**.

Before we can understand the why and the evolution of reading aspects we must first look at the how. One example of this **how** lies with another concept known as **network theory**.

Definition by Moretti: "This is a theory that studies connections within large groups of objects: the objects can be just about anything. and are usually called nodes or vertices; their connections are usually called edges." (Moretti, 222).

example 1

After giving this definition, Moretti goes on to begin discussing an example of distant reading created using **Shakespeare's Hamlet and network theory**.

The image shown on this slide used the characters of Hamlet as what Moretti referred to as **nodes or vertices**, and their **interactions as the edges**.

However, as Moretti himself admits, in this image characters who exchange a single line, and those who exchange hundreds are treated the same in that the edges only represent that an interaction was had by the two characters, nothing more, and nothing less. This makes **the characters simply points of data connected by the lines of interaction on a mind map.**

The reason that the image is **considered a distant reading** of Hamlet is because it **simply presents a series of data points** that **can then be interpreted in various different ways**.

One such way that Moretti **chose to interpret the data was by removing Hamlet** from the image to see what would happen.

As the new example shows, not much changes except that Horatio now becomes the only person bringing both the left and right sides of the image together. By removing Hamlet from the diagram Moretti was able to prove how central Hamlet is to the plot of the story without having to spend a vast amount of time looking at and analyzing the content of the character's dialogue and actions.

example 2 - The rise of the novel

The previous definition of distant reading was that it is used to analyze more than the content and structure of a few books. However, the Hamlet example shows that it can also be applied not simply to books, but the cast of characters within them.

Another example from Moretti's Graphs, Maps, Trees, can be seen to have done a distant reading of the rise of the novel from the eighteenth to the twentieth century. The second example may be talking about the rise of the novel, but it's **not "reading" actual literature**. Yet, if a timeline can be "read" then what is reading? How does this new form of reading operate?

The new version of "reading," distant reading came about as a result of the **advancement of digital technologies**.

"The width of the corpus and the speed of the search have increased beyond all expectations: today, **we can replicate in a few minutes** investigations that took months and years of work (Moretti, 221).

As Moretti states in this quote **digital technology has made it easier to research all kinds of data**, thus making it easier to analyze literature, however that is not all it does.

Some other notable examples

Ted Underwood's work:

The Work of Ted Underwood Ted Underwood has been among a pioneering group of scholars who analyze the textual data in large digital collections to figure out how to answer bigger picture questions. As a **scholar of English literature**, he uses digital technologies to try to get a broader look at **how literature has developed over time**. One technique he has used is to apply computational methods to distant reading, an approach that allows an investigator to evaluate the content of many, possibly even one thousand or more, books at a time with the assistance of digital tools.

One area Underwood has focused on in **his recent work is gender**. Along with **David Bamman and Sabrina Lee**, he assessed a corpus of texts, including those in the **HathiTrust**, to find out what it could **reveal about gender and authorship in English-language fiction texts and how gender roles informed characterization in those texts**.

The group’s conclusions suggest that characterization of gender has been less rigidly defined over time, but the **percentage of authors who identify as women and the percentage of characters identified as women has decreased over time**. As they put it, **“While gender roles were becoming more flexible, the space actually allotted to (real, and fictional) women on the shelves of libraries was contracting sharply.**”

Using data to spot trends like this helps us reflect on what those trends say about our society and such work opens the door for scholars of disciplines such as history, literary studies, or gender studies to ask new questions.

reading beyond text

With the advent of digital technology a much larger quantity of literature has become available to us and it is not always in the form of text. For example, Lev Manovich and seven others have created a website based around a relatively new form of literature known as the selfie. How is the selfie a piece of literature you may be asking? Well that al depends on how it's "read"!

A link to Selfie City can be found here: http://selfiecity.net/

The way Lev and his colleagues chose to "read" the selfie was by gathering data based around selfies taken in five different cities: New York, Sao Paulo, Bangkok, Moscow, and Berlin. The data they compiled was based around camera angle, gender, age, and mood amongst other factors.

The way they read this data they had gathered from Instagram, was to create graphs and charts as visual representations of the data that could then be toggled to focus in on different aspects. For example, only female selfies from New York could be focused on while the rest of the data in the categories of gender and location are ignored. This would then change what selfies are displayed at the bottom of the screen as if one had made a google images search with the same parameters.

This is considered distant reading because it is analyzing a large amount of data to show how different external factors affect the data.

The research involved in creating readings such as these has, as Moretti would argue, gotten easier thanks to digital technology. However, it has also gotten easier to create statistics based on the data such research brings to light now that one can simply plug bits of data into the proper digital tool and let thecomputer do the rest.

The questions

Having gone over what distant reading is, how it differs from close reading, and having briefly touched on some examples of how distant reading was made to work with digital technology, and analyze both text based literature and images, it is time to address the questions: what is literature and what is reading?

1. What is literature?

In the past when one thought of the word literature they thought of novels or scholarly text books. Even the Oxford English Dictionary's top result is to define literature as "Familiarity with letters or books; knowledge acquired from reading or studying books, esp. the principal classical texts associated with humane learning." (OED Online, Literature, n Definition 1) However, this is no longer the one and only true definition.

Distant reading's nature of reading and analyzing data sets such as paintings grouped by color has created a much broader version of literature's definition than "knowledge acquired from reading... books" so that now books are only a fraction of what it is we refer to as literature.

As seen in the Hamlet mind maps, the characters and the concept of who interacts with whom has in and of itself become literature. The number of novels created across two whole centuries is now literature. Even selfies taken in New York have become literature.

2. What is reading?

Turning to the OED once again it can be seen that the commonly accepted definition of reading is "to understand what is meant by the letters or signs." So from this we can infer it to be referring to text-based printed books, like text books novels, or poetry collections, and not much else.

This may seem like a fallacy based solely upon one's ability to read things like comic books which mix text and images. However, the fallacy extends beyond that if we can read selfies and the numerical values of timelines.

Some may think that this is why reading has now separated into the two types known by the monikers "close" and "distant." One for reading books, and another for reading other things.

Considering Moretti created the concept of distant reading as a possible way to read the books which fall under his concept of "world literature," the assumption that the two types of reading are meant for two different types of literature is only half right.

The assumption that the two types of reading are meant for two different types of literature is only half right. This means that it is also half wrong. This is because as previously stated:

1. both types of reading can be applied to books,
2. and, also, what hasn't been discussed is that close reading can likewise be applied to non-text based literature

An example of this can be found by looking at Moretti. As we've already seen he removes Hamlet from the character diagram of Hamlet as a way to show that Hamlet is one of the central figures of the play. However, he also removes Horatio at a later point in his discussion of network theory and is surprised to see that Horatio and Hamlet, if both removed, completely separate the characters on the left side of the chart from those on the right. He himself also points out that "the Ghost and Fortinbras- which is to say, the beginning and the ending of the play- are completely severed from each other and from the rest of the plot! (Moretti. 231).

Here we can see **the beginnings of a close reading of the Hamlet network** based upon something as simple as removing two characters from it.

So if the image of a network can be close read, and the plot of Hamlet can be distant read, thus creating said image, what does this say about the concept we refer to as "reading"? Think about it for a moment. It tells us that what we can and cannot read has changed and how we close read, and now distant read, has adapted to include new forms of media and therefore like literature requires a new definition.

Conclusion

When Franco Moretti first coined the term 'distant reading' in 2000, he used it with a meaning reminiscent of the compilatory origins of the concept, similar to "second-hand reading": using research literature, metadata or other short-cuts like titles and subtitles instead of reading the full text. From this starting point, and in parallel with more computational and more quantitative practices, Distant Reading has evolved to designate any computational, but especially quantitative, method of literary text analysis - so much so that the term now 'self-evidently implies computation' (Goldstone 2017, 637; see also Underwood 2017 and Bode 2017).

A fundamental assumption of the earlier concept of 'distant reading' was that because metadata or secondary literature are created by humans who have read the full texts, they can stand in for the full text. Also, that the bird's eye's view provides insight into the longue durée and into literature as a system (Oberhelman 2015). A fundamental assumption of current Distant Reading research is that useful (even if imperfect) formal and quantifiable textual features can be used as indicators or proxies for relevant literary phenomena, hence the centrality of modeling (see McCarty 2005; Flanders and Jannidis 2019) in Distant Reading research practice. Finally, the idea that despite the broadening meaning of the term “literature” (decanonization), literary texts have a specific way of functioning that requires the adaptation of methods to this domain.

A critical framework for discussion of results of any data mining and text analysis is required to keep it from standing on its own, decontextualized. While patterns emerge, and large trends can be discerned, the question of what these are indicative of remains. Do they only show trends in the data? Or can they reveal trends in phenomena of the actual and lived world (Lee 2019)? What is the relation between these two? Abstraction, extraction, reduction, and simplification are frequently used terms in discussing large data sets. The value of the research needs to be measured against these considerations. As in so many aspects of digital work, the value is in the dialogue with traditional methods.

- Wout Dillen's presentation
- Ryan's presentation on DH work that is interesting
- Ted Underwood's work
- Introductions of Distant Reading publications
- The Computational Case against Computational Literary Studies
- Khatib, workflow of distant reading

Historical roots

The pre-history to the concept now covered by the term 'distant reading' reaches back to the 15th century, when a rhetorical topos of "too many books" appeared (see Blair 2011). The solution was in excerpts and encyclopedias, based on the principles of compilation and summarization. The goal was to provide access to the essence of all relevant books instead of having to see them all at the same time. Of course, quantitative approaches to literary texts have appeared before the advent of computing (e.g. Mendenhall 1887) and computational approaches have diversified before the term 'distant reading' appeared (e.g. Ellegård 1962, Mosteller and Wallace 1963, Burrows 1987; see Hockey 2000).

Some notable cases

Moretti, Franco. *Graphs, Maps, Trees: Abstract Models for Literary History* (2005) Jockers, Matthew. *Macroanalysis: Digital Methods and Literary History* (2013) Underwood, Ted. *Distant Horizons: Digital Evidence and Literary Change* (2019)

Dissidents

Stephen Marche, "Literature is not Data: Against Digital Humanities" (2012) Nan Z. Da, "The Computational Case against Computational Literary Studies" (2019)

Main argument of Nan Z. Da: She argues that the methods used in the field are not reliable and that the results are not replicable. Moreover, she argues that the field is not able to produce new knowledge. Some examples she gives for this are the following: Underwood's "The Life Cycles of Genres" (2017) and Jockers' "The LDA Buffet is Now Open; or, Latent Dirichlet Allocation for English Majors" (2013).

What's behind it?

- Modeling
- nlp
- confirmation bias
- bag of words

The way to go (workflow)

Discussion about methodology in the humanities

References

Arnold, Taylor, and Lauren Tilton. 2019. “Distant Viewing: Analysing Large Visual Corpora.” Digital Scholarship in the Humanities. https://doi.org/10.1093/digitalsh/fqz013.

Blair, Ann M. 2011. Too Much to Know: Managing Scholarly Information before the Modern Age. New Haven: Yale University Press.

Bode, Katherine. 2017. “The Equivalence of ‘Close’ and ‘Distant’ Reading; or, Toward a New Object for Data-Rich Literary History.” Modern Languages Quarterly 78 (1): 77–106. https://doi.org/10.1215/00267929-3699787.

Burrows, John. 1987. Computation into Criticism: A Study of Jane Austen’s Novels and an Experiment in Method. Oxford: Clarendon Press.

Da, Nan Z. 2019. “The Computational Case Agains Computational Literary Studies.” Critical Inquiry.

Ellegård, Alvar. 1962. A Statistical Method for Determining Authorship: The Junius Letters, 1769-1772. Gothenburg: University of Gothenburg.

Flanders, Julia, and Fotis Jannidis, eds. 2019. “The Shape of Data in the Digital Humanities: Modeling Texts and Text-Based Resources.” Digital Research in the Arts and Humanities. London & New York: Routledge.

Goldstone, Andrew. 2017. “The Doxa of Reading.” PMLA.

Hockey, Susan. 2000. Electronic Texts in the Humanities: Principles and Practice. Oxford: Oxford University Press.

McCarty, Willard. 2005. Humanities Computing. New York: Palgrave Macmillan.

Mendenhall, Thomas C. 1887. “The Characteristic Curves of Composition.” Science ns9 (2145): 237–46.

Moretti, Franco. 2000. “Conjectures on World Literature.” New Left Review, no. 1.

Mosteller, Frederick, and David L. Wallace. 1963. “Inference in an Authorship Problem.” Journal of the American Statistical Association 58 (302): 275–309.

Mueller, Martin. 2012. “Scalable Reading.” 2012. https://scalablereading.northwestern.edu/?page_id=22.

Oberhelman, David D. 2015. “Distant Reading, Computational Stylistics, and Corpus Linguistics: The Critical Theory of Digital Humanities for Literature Subject Librarians.” Digital Humanities in the Library: Challenges and Opportunities for Subject Specialists. Chicago: Illinois: Association of College, Research Libraries. https://shareok.org/handle/11244/33193?show=full.

“On Franco Moretti’s Distant Reading.” 2017. Publications of the Modern Language Association (PMLA) 132 (3): 613–89.

Schulz, Kathryn. 2011. “The Mechanic Muse - What Is Distant Reading?” The New York Times. https://www.nytimes.com/2011/06/26/books/review/the-mechanic-muse-what-is-distant-reading.html.

Underwood, Ted. 2017. “A Genealogy of Distant Reading.” Digital Humanities Quarterly 11 (2). http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html.

Underwood, Ted. 2019. Distant Horizons: Digital Evidence and Literary Change. Chicago: The University of Chicago Press.

Calit2ube. "Cultural Analytics - Mark Rothko Paintings - on the 287 Megapixel HIPerSpace Wal at Calit2." Online Video Clip. YouTube. YouTube, 8 July 2009. Web. 16 December 2014.

OED Online. Oxford University Press, December 2014. Web. 17 December 2014.

Manovich, Lev. Selfiecity. The Graduate Center, City University of New York, California Institute for Telecommunication and Information, and The Andrew W. Mellon Foundation, 2014. Web. 16 December 2014.

Moretti, Franco. Distant Reading. Google Books. 2013 Web. 12 December 2014 "reading, n." OED Online. Oxford University Press, December 2014 Web. 17 December 2014.

Schulz, Kathryn. "What is Distant Reading! New York Times. 24 June 2011. Web. 15 December 2014.