**What Was Distant Reading?**

Wednesday October 14, 4:30 pm

Slides: theotherdh.com/presentations/DistantReading

Intro: thanks to Paul and Jeremy for inviting me, and thanks to Paula and Riva for such a warm welcome.

Today I am going to talk about work Laura Heffernan and I are doing on three women distant readers from the first half of the twentieth century: Edith Rickert, Caroline Spurgeon, and Josephine Miles. And I’m going to ask what we can learn from them for work now, what present and future of distant reading a past that includes them might point to.

This work comes from a book project Laura and I are finishing called The Teaching Archive: A New History of Literary Study. In the book we rewrite the history of literary study from the perspective of classrooms at a wide range of higher ed institutions to show how the real history of literary study before 1970 has a much richer account of method and a far, far wider canon than we currently imagine.

The women distant readers I’ll be talking about today kind of snuck up on us; we weren’t looking for them, and we know there are many more of them than we currently have studied. So now we’re trying to pay some more attention to them, and think through the consequences of what we’ve learned about their work for thinking about “distant reading” – quantitative text analysis – today.

(Chat: when Laura and I decided that we needed a place to start testing this work and that this would be an ideal place it was already too late for her to fit this into her schedule, but I just want to be clear that this work is entirely by the two of us. And in fact I’ll talk about collaborative work of various kinds IN this talk , and I’m also happy to talk about the history and practice of collaborative work in q and a.)

[mention slide location, ask them please not to share or circulate]

So before I talk about its history – what is distant reading today? We usually use the term to talk about the processing of information about or drawn from a (relatively) large number of texts, and we usually use the phrase in the domain of the analysis of unstructured texts we want to look at for the purpose of literary or cultural analysis. “Distant reading” implies literary study, of course, because as a metaphor it opposes itself to “close reading.” But the opposition is – as many people have pointed out ad nauseam – merely rhetorical. From a methodological perspective, it doesn’t make sense to talk about a choice between the two, and there’s no necessary contradiction in using both. (Probably everyone who came to a talk called “What Was Distant Reading” already knows this, but just in case.)

It’s also worth saying that distant reading is inherently quantitative – it turns texts into numbers – but it is not necessarily computational. Ted Underwood notes in his 2017 article “A Genealogy of Distant Reading” that the central practice that distinguishes distant reading is not computation or the use of “big data”, but “the practice of framing historical inquiry as an experiment, using hypotheses and samples (of texts or other social evidence) that are defined before the writer settles on a conclusion.” (For Ted specifically, the promise of distant reading today is about the way it draws on recent innovations in statistics that have overcome “the difficulty of representing unstructured text,” allowing “researchers to include qualitative evidence like text in a quantitative model by the simple expedient of using lots of variables (say, one for each word).”)

Distant reading also distinctively requires what Sarah Allison calls “reductive reading,” the choosing of a quantifiable feature or features of a text to stand for the text itself; we might, in Stephen Ramsay’s formulation, even say that when we turn texts into data we are not reducing them, but actually transforming them into something else. And, crucially, what those reductions or tranformations of literary texts into information allows us to do is to compare larger numbers of texts with more precision and - often with more nuance - than we usually are able to do.

The phrase “distant reading” itself IS new - it was coined by Franco Moretti a few decades ago, and has come to be widely used as a term for current work in the quantitative analysis of literary texts. But this TYPE of work is anything but new; various critics have noted distant reading’s origins in 1950s literary sociology (as Ted Underwood does) or in 1960s computer-assisted humanities work; many cite Roberto Busa’s IBM machine-aided concordance to the work of Thomas Aquinas as an origin point for specifically *computational* text analysis.

But even these mid-twentieth-century historical origins points don’t go back far enough.

Before the mid-twentieth-century, distant reading was the province of women scholars. Their quantitative methods did sometimes seemed new to them, but they were also closely connected to the traditional literary scholarly practices of concordancing, indexing, and scholarly editing. And so these early distant readers offer us a history of quantitative text analysis as much more continuous with the history of literary study than the current language of newness, computation, and innovation that often accompanies distant reading. (Recent distant reading in fact HAS these same dependencies; it often relies heavily on bibliographies, but its practitioners don’t emphasize this.)

We also find that this early distant reading work was also avowedly and deeply collaborative in the way it was created, and collective and in the way it imagined future scholars using it; it was often practiced as or alongside students in classrooms and labs, it stressed the crediting of collaborators, and it emphasized both the interpretation of data and the sharing of processes and datasets with others who might build on it or interpret it differently.

And matching this sense of the work’s process, we find also a growing interest in the project of using quantitative methods to discover the collective language of literary production, the sense of a shared project that aims to reconstruct what Josephine Miles called the “common materials” of literary language, and to compare the ways in which different writers used their shared resources.

So now I’m going to talk briefly in turn about Edith Rickert, Caroline Spurgeon, and Josephine Miles. Rickert and Spurgeon, almost identical contemporaries working in the 20s and 30s, seem unaware of each other’s work as distant readers; Miles, whose works spans the late 30s through the 70s, cites the quantitative work of both Spurgeon and Rickert.]

3. Edith Rickert (1871-1938)

Edith Rickert was professor in the English department at the University of Chicago. She had done her graduate work there at the beginning of the century, and then spent some time teaching high school on the east coast and living in London making a living writing short stories and non-fiction for journals and magazines before returning to the US.

Rickert’s text analysis work focuses on the way the quantitative analysis of individual works or small groups of works could help readers develop a more concrete sense of a literary work. The methods that Rickert and her students developed worked by defining specific textual characteristics for intensive study and using counts and visualizations as an aid to understanding those characteristics. This “mechanical” (as she termed it) approach to literary text worked by developing a set of analytic protocols or practices to aid individual readers’ understanding and appreciation of literature.

[TITLE PAGE]

Her 1927 book New Methods for the Study of Literature records the classroom-based experiments in this kind of quantitative analysis of literature that she performed with her graduate students in English 376: Scientific Analysis of Style. In English 376, Rickert invited her students to help her devise a series of methods for comparing and visualizing literary texts by quantifying and visualizing aspects of their form and language -- images, sentence-lengths, word choices, rhythms, and connective phrases.

As she explained in the introduction to New Methods, these experiments drew upon “the methods of code analysis used in the Code and Cipher Section of the Military Intelligence in Washington” she had learned while working there during World War I alongside the other academic philologists (including her mentor University of Chicago professor John Manley) who assisted with code-breaking during the war. (Continued at Riverbank cryptography, etc.)

IMAGE

The beginning of the book offers readers experiments in the simple taxonomizing of imagery – readers might use the taxonomy to mark up individual poems on notecards; they might then count the proportions of image-bearing words in a poem or group of poems,

IMAGE

and visualize the proportions in a chart.

IMAGE

To study “thought-patterns,” readers might choose a simple metric like number of syllables per sentence, and count a sample of 100 or 1000 sentences, graphing them to prepare for comparison between poems or authors.

IMAGE

To emphasize line length on a page, one can also use simple but striking technique for reducing a poem or prose page to nothing but its title and line lengths:

IMAGE

Through the building up of these kinds of analyses of sound, charts, summaries, and comparisons, Rickert notes, “differences which do not appear in reading can be observed” (219).

Through these methods, Rickert and her collaborators attempt to train sensitivity to literature through allowing the reader to share “in the very process of creative thought” through minute attention to language, building from the word to the sentence to the image.  If the “subtle effects” of language can be “traced to their sources,” Rickert wrote, we might, “through mechanical means at first” be able “to train the attention to look for the definite qualities into which literature can be analyzed, to recognize certain effects as due to rhythm, others to sound, others to imagery, others to the emotional associations of words, others to the very architecture of the thought, or even to the placing of the visible words on the page?” This granular analysis of poetry could provide the basis for a shared consensus, at least, of the basis upon which a literary texts mean is made.

She was aware that the methods are “so new that they are valuable mainly for opening up the way to further study of this kind.” The particular classifications may or may not be the best – something which must be “proved or disproved by further experiment.” Yet the charts and summaries in the book - which, Rickert emphasizes, are not single workings out of the lines by single people, but charts and summaries that have been made “over and over again” – there is the beginning of the basis for an interpretation of literary language based on a consensus rendering of the poem.

The iterative and experimental nature of this work was crucial to Rickert. She wrote that “the educational value of the work lies largely in the fact that students become keenly interested in it and are constantly devising and suggesting new and better methods.” The footnotes in New Methods memorialize this classroom collaboration by diligently recording the names and “suggestions of the students.”

[Worth pointing out about experimental and collaborative literary study in the 1920s, which Laura and I talk about in our book; Dewey, etc; K-12. The collaborative aspects of *New Methods* were also part of a reconceptualization of collaborative humanities work as new. In 1924, Stanley Greenlaw presented to the Association of American Research Universities on “Recent Movements for Co-operative Research in the Humanities,” arguing that humanities researchers needed organizations and funding to co-ordinate research projects and share materials just as much as scientists did; while the handful of large research universities where graduate education went on already had these facilities, he noted, the research capabilities of PhDs taking jobs at other kinds of institution was hampered by lack of coordination.]

**4. Caroline Spurgeon (1869-1942)**

Caroline Spurgeon was the first female professor in the UK and author of Shakespeare’s Imagery and What It Tells Us (1935). Born in India in 1869, and orphaned by the time she was ten, she spent some of her early childhood in Germany and France before being sent to school in England. She began her university-level study at Kings College and University College, London, going on to do pass university examinations in English Literature and Language at Oxford - but not, because of her gender, to receive a degree - in 1899; she ultimately received her doctorate from the Sorbonne based on her work the history of literary criticism of Chaucer (she’s originally wanted to work on George Eliot). She had a remarkable, wide-ranging career; she was a member of the Newbolt committee, contributing to their report on The Teaching of English in England. And with longtime companion Virginia Gildersleeve (for many years the Dean of Barnard College) and other collaborators she founded the International Federation of University Women.

BOOK COVER IMAGE

Like Rickert, Spurgeon believed that mechanical, informational approaches to literary text could circumvent preconceptions and unlock unexpected patterns. Shakespeare’s Imagery and What it Tells Us, her most famous and last-published major scholarly work, focuses on undoing the fixed and finished form of the canon. Published after a decade of work “gradually selecting, sorting, and classifying” all of the metaphorical uses of language is Shakespeare’s corpus using notecards, the goal was to mine Shakespeare’s work for “neutral content,” - or, as Spurgeon put it “their stuff.” As she explains, “I treat Shakespeare’s images as documents.” Spurgeon snips the link between tenor and vehicle, discarding the former and cataloging the latter. So in “they’ll take suggestion as a cat laps milk,” she discards “they”and “suggestion” - the fiction’s referents, as it were - and keeps and catalogs “cat” and “milk.” For Spurgeon, this indexed data was as important as her book; Spurgeon hoped “eventually to publish the material itself, so that other students can check and perhaps extend it, in order that it may serve as data and a starting-point for further research of various kinds.” She didn’t, because she died soon after publishing Shakespeare’s Imagery, but the notecards are in the Folger, a digitization project waiting to happen.

So in Shakespeare’s Imagery Spurgeon - oddly - creates the raw material for a totally different mapping of Shakespeare’s works out of the non-referential aspects of his fictional worlds - the metaphors - and converts them back into referential content. Rereading the vehicles of Shakespeare’s metaphors as real worldly things that originated in the everyday experience of the authors who use them, she unravels the tightly-knit formal unit of the metaphor (or “image” in her language) to return its component parts to the world of the not-yet-literary. Turning her attention to what one might think of as the most “literary” part of of the works of the most canonical English writer in history, she in effect undoes their literariness by returning to figural language’s origins to the everyday world.

As Shakespeare’s Imagery unfolds, the really central point becomes this slow decomposition of Shakespeare’s finished and canonical works into an index of things that interested Elizabethan playwrights, things they noticed and eventually made into literature - from Shakespeare’s detailed observations of birds and his preference for green vegetables to Bacon’s attraction to the world of learning to Marlowe’s preference for city over country.

IMAGE

This new method offered a more objective view of the plays, Spurgeon suggested, because it bypassed the evaluative and canonizing moves of traditional criticism through indiscriminately collecting together ALL of Shakespeare’s uses of figural language. While many critics had already examined selected Shakespearean images from across his corpus, in her preface Spurgeon explains that “the novelty of the procedure I am describing is that ALL his images are assembled, sorted, and examined on a systematic basis, the good with the bad, the disagreeable with the unpleasant, the coarse with the refined, the attractive with the unattractive, and the poetical with the unpoetical.” Not selected to illustrate a “preconceived idea or thesis,” “they are studied, either as a whole, or in groups, with a perfectly open mind, to see what information they yield, and the result comes often as a complete surprise to the investigator.”

In a section of a chapter on “Shakespeare’s Tastes and Interests,” for example, Spurgeon argues that Shakespeare’s “observation of and interest in needlework” can be concluded from the very specific knowledge displayed in figural references to “a bodkin, a silken thread, a twist of rotten silk, or an ‘immaterial’ skein of sleave or floss silk and needles, threaded and unthreaded” in plays including Love’s Labours Lost, Much Ado, Coriolanus, and King Lear. She notes that the range of Shakespeare’s domestic references indicate is familiarity with very modest domestic setting, especially in comparison with the more opulent references to domestic settings she finds in plays by Thomas Dekker, and tracks the way Shakespeare’s references to food and become more precise while exponentially expanding in number in works he wrote after 1594.

But more is at stake in this conceptual flattening of tenor and vehicle than a claim to objective literary analysis - a goal which actually recedes as the book unfolds across fifteen chapters, eight appendices, an extensive index, and seven charts with interpretive text. Throughout SI, Spurgeon pulls away from the sense of any play as an organic whole or a finished work, instead using Shakespeare’s writing to generate an imaginative map of the earliest moments of its composition, the moments Roland Barthes theorizes as the author’s jotting down of brief “notula” as they occur to him in the moment. **Her criticism seeks to reveal literature as living rather than canonical by thinking about how literature might be composed from a variety of the author’s referential, iterative, everyday experiences.** What emerges is Spurgeon’s emphasis on the decomposition of Shakespeare’s complete works into a dataset that lets us study everyday life in Elizabethan England, and Shakespeare’s everyday life in particular. We can of course learn about Shakespeare indirectly from his finished plays, she notes, but we can also “add a little to our detailed knowledge of Shakespeare as a person by studying the data he has left us, incidentally and by the way as it were, embedded in his images." Abandoning the conventional sources and method of historicist work that had been long familiar to her, Spurgeon’s turn to the plays themselves give new form to her longstanding interest in the details of literary composition and the process of canon formation.

[Spurgeon’s bypassing of the evaluative and canonizing moves of criticism rubbed some critics the wrong way. Many of her reviewers suggested that Shakespeare’s Imagery was more a concordance than a work of criticism; review after review framed Spurgeon as an information worker whose writing could nevertheless profitably used as raw material by more brilliant (usually male) critics. Stanley Hyman was typical of many critics when he complained that her scholarship was the result not of “superior insight but merely through her opening up a new area of information, the image-tabulations” and suggested that her book was a reference tool superficially disguised as a scholarly monograph.]

But some contemporaries saw a different kind of value in SI. Reviewing the book in 1944, Rachel Hannon wrote that Spurgeon’s work held out the alluringly promise that "by an almost mechanical means we baser mortals can come upon a grasp of poetic values instantly apparent to Sidney, Shakespeare, Shelley, or Coleridge.” For Hannon, the “almost mechanical” nature of Shakespeare’s Imagery demystifies the creation of “poetic values” in a way that lets everyday people connect to and understand it. Using “almost mechanical” methods to dismantle their finished forms back into their component parts, Spurgeon’s distant reading shows how major works and canonical reputations are composed out of references to everyday observations.

**5.Josephine Miles (1911-1985)**

Josephine Miles – like Spurgeon, whose work she cited – was also interested in decomposing finished, canonical literary works back into their component elements; from her early thesis work on Wordsworth’s adverbs to her mid-career work on nouns, adjectives, and verbs in five centuries of English poetry and prose to her later interested in concordances and connectives, Miles pursued the quantitative measurement of continuities and changes in literary language.

After receiving her BA from UCLA and her PhD from Berkeley, Miles worked at Berkeley from 1939 through the 1970s. Miles had an amazingly long and prolific career; she is remembered as an important poet and a remarkable teacher. She is much less well remembered as a prolific and pioneering scholar of quantitative and computational text analysis in English, despite having published dozens of monographs and books drawing on her datasets, as well as overseeing the first-published computationally assisted concordance.

Miles’s first distant reading project was her graduate thesis -- an analysis of the adjectives favored by Romantic poets, published in 1942 as Wordsworth and the Vocabulary of Emotion; in the same year, she published The Pathetic Fallacy in the Nineteenth Century, an extension of her work with adjectives across the nineteenth and into the early twentieth century. In the 1940s, with the aid of a Guggenheim grant she expanded this work into a large-scale study of the phrasal forms of the poetry of the 1640s, 1740s, and 1840s, published in 1946 as Major Adjectives in English Poetry From Wyatt to Auden. In the following decade she continued to expand her dataset and published books based on it, including The Continuity of Poetic Language (1951), Eras and Modes in English Poetry (1957), and Renaissance, Eighteenth-Century, and Modern Language in English Poetry: A Tabular View (1960).

Like Spurgeon, Miles’s quantitative work came from her interest in where poets got their language. But while Rickert and Spurgeon’s interested weighted more towards an interest what made poets distinctive from one another, Miles’s work emphasized continuity; Miles was interested in individual style, but tended to defer such questions in favor of seeking the common language of poetry in different eras. As she wrote in the beginning of her 1948 monograph “The Primary Language of Poetry in the 1640s,” her goals was “to describe the common materials of language through which poets work is to describe at once the limitations and the potentialities of their medium” in order to develop “a theory of major language” in each era.

Her approach, she wrote, made “compromise between individual and society by its focus on the contemporary agreements in the practice of individual poets”; by looking at the major language of poets in a decade or era, “one may attempt to define some of the mass, the density, of poetry as a human production in which form celebrates value and makes it most weighty.”

Miles’s quantitative work was an even longer-term project than Spurgeon’s. While Spurgeon and her assistants took a decade the produce the data about early modern playwrights’ images, Miles began working on her word counts in the late 1930s, and continued them for decades; they were created by hand, with the help of generations of graduate students. She began by examining the poets of the 40s in each century, beginning from the 1640s and working through the 1940s (before returning to the 1540s after her early modernist colleagues complained). Drawing on Jyotish Chandra Ghosh’s listing of poets in his Annals of English Literature, Miles would take an 1000-line sample of twenty poets and tabulate the nouns, adjective, and verbs. She notes that she considers her 1000-line amount as “provisionally’ “representative” but that “I therefore mean to make no statement about any more than the lines under consideration.”

IMAGE

In an oral interview in the 1970s Miles remembers what the everyday work of this kind of data creation was like:

Then I developed a whole theory that I wanted to follow out, the one of major language. Then I just had a built-in job to do every summer. I sat on the patio with my little beat-up traveling typewriter that had only three banks of keys, and typed out these studies of the language of the poets of the 1640s, 1740s, and so on. These were good for getting - I wanted something too, not only that I liked to do, that I could sit outside with and enjoy, but also that students could help with so that they could get support grant money and this kind of analysis of language they could do. Over the years, in fear and trembling, I’ve gone back and checked whether they did it well, and on the whole I’m sure there are terrible errors still but on the whole they did beautifully. They were responsible and good people. I just did that until about 1951, I guess; for about a decade I did that, and then additions through the sixties.

Miles’s results included lists of both “major” vocabulary – words used more than ten times by a majority of the poets – as well as more “minor” and individual uses of all of the most frequent tabulated words.

[IMAGES – slide through hand and print]

She would also calculate the noun-adjective-verb proportions. As she says, errors are always possible but “most of the major terms are here, in their proportions and proportionate contexts, and their principle is here, the principle of quantitative emphasis as clue to qualitative whole, as this study has proposed it.”

She is aware that “To many readers such analysis of quantity and connection will seem alien to the modes of poetic creation.” But, she asks, “is it not possible that proportions once poetically felt, and then numerically discovered, may be poetically re-perceived?”

While Miles and her students created all of the data for this long-term project by hand, in the early 1950s Miles also became responsible for remaking abandoned Dryden Concordance, an index card based project that a former colleague had been working on. Miles contacted Berkeley’s Cory Hall computing lab and asked them if they thought they could help; they could. She ended up working with, among others, Penny Gee, a Cory Hall computer operator who was one of the many often unremembered women working in computing at the time.

I’m not going to talk in detail about the Dryden concordance project since Laura and I have published an essay on it and I’ll give you a link; it’s worth pointing out that that as far as we know Miles and her collaborators were the earliest people to use IBM machines to create a concordance; Miles they were cited by the Cornell Concordance team. (We’re interested in the Cornell Concordances, doing some work in the archives.)

I want to point out, though, the way thinking about machine concordancing influenced Miles’s thinking about the quantitative analysis of literature. In a 1965 review of some new concordances, Miles saw the distant promise of computational analysis’s ability to represent unstructured text with great flexibility and granularity; as she wrote

“But now it seems to me that the new possibilities all lead toward inclusiveness: not so much of reproduced text as of analyzable and quantifiable elements, so that more and more we shall be able to learn about underlying structures of connection and repetition within one text and from one text to another.”

This interest was matched, in her own work, with a turn – still gathering data by hand – to the kinds of very common words that were left out of concordances and word counting early on – particle words like connectives and pronouns whose work of replacing revealed significant patterns. So Miles’s career, in spanning hand-made tabular data and computational text analysis work, also offers a window into the ways that computational **thinking**, not computational power, influenced literary study; Laura and I are trying to think this through further.

So to wrap up, what can we learn from the work of Rickert, Spurgeon, and Miles for thinking about distant reading now?

Of course there’s the part about a more accurate and more inclusive and more interesting history, that makes distant reading seem more accessible to more people, possibly.

And there’s a history of computation piece – the more of this kind of work we do, the more it seems like collaborations between literary scholars and linguists and computer scientists have only recently come to seem unusual.

There’s also the sense of textual closeness that one gets from this kind of quantitative, non-computational work that is actually useful for both doing and teaching quantitative work – the starting with an individual text and building one’s way up, using pen and pencil at first.

There’s a strong idea that quantitative work is not opposed to literature or everyday reading or literariness, but a part of understanding the language that we all share when we begin the work of composition, whether the composition of a first-year essay or an accomplished writer’s poem.

There’s a strong sense – stronger than in current distant reading work, I think, despite calls to standardize and shared data – that the work of people like Rickert, Spurgeon, and Miles was designed to be collaborative and oriented towards re-use and the sharing of data as well as of results.

And there’s also, especially in Miles, a sense that the affordances of computation – the ways computational processes can help us think about the construction of literary texts – can help us not just in practice, but as theory.