CHAPTER 3

Modeling Modernist Dialogism: Close Reading with Big Data

Adam Hammond, Julian Brooke, and Graeme Hirst

Of the many bold pronouncements in Matthew Jockers's groundbreaking *Macroanalysis* (2013), perhaps the boldest is his claim that the advent of computational "distant reading" will make close reading obsolete as a method for investigating literary history. Jockers argues that the development of massive digital literary corpora has placed literary historians in a position in which they no longer need to rely on "partial sample" close readings, but can instead perform "investigations at a scale that reaches […] a point of being comprehensive."[1] Jockers writes:

This work was financially supported in part by the Natural Sciences and Engineering Research Council of Canada and the Social Sciences and Humanities Research Council of Canada.

A. Hammond (✉)
San Diego State University, San Diego, CA, USA
email: ahammond@mail.sdsu.edu

J. Brooke
University of Melbourne, Melbourne, VIC, Australia

G. Hirst
University of Toronto, Toronto, ON, Canada

S. Ross, J. O'Sullivan (eds.), *Reading Modernism with Machines*,
DOI 10.1057/978-1-137-59569-0_3

> Science has welcomed big data and scaled its methods accordingly. With a huge amount of digital-textual data, we must do the same. Close reading is not only impractical as a means of evidence gathering in the digital library, but big data render it totally inappropriate as a method of studying literary history.[2]

Before Jockers, even the most passionate defenders of computational literary analysis tended to stop well short of the language of the "totally inappropriate." Though Susan Hockey champions the "rigor and systematic unambiguous procedural methodologies" of computational analysis against the "serendipitous" procedure of close reading[3]—and although Julia Flanders argues that the computer is not just a "substantiator" of close readings but indeed "a device that extends the range of our perceptions to phenomena too minutely disseminated for our ordinary reading"[4]—neither goes so far as to claim that computational analysis could, or should, replace close reading. Stephen Ramsay and Tanya Clement, digital humanists responsible for perhaps the most illuminating digital work to date on modernist texts, are more careful still in insisting on a symbiotic relationship between close reading and computational analysis. In his work on Virginia Woolf's *The Waves*, Ramsay presents quantitative analysis as a digital-age method of achieving the Russian Formalists' goal of *ostranenie*—a making-strange that clears the path for renewed close reading.[5] Clement likewise presents her computational work on Gertrude Stein's *The Making of Americans* as a way to "defamiliarize texts, making them unrecognizable in a way (putting them at a distance) that helps scholars identify features they might not otherwise have seen."[6] For Clement and Ramsay, computational analysis can thrive only in an ecosystem of close reading, since its proper role is to enrich existing close readings and to prompt further ones.

Despite Clement and Ramsay's shared vision of a mutual cooperation between human and computational reading—and although Clement uses the term "distance"—what they advocate is not a combination of close and distant reading per se. While both seek literary insight through quantitative analysis, neither employs big data approaches or works at Jockers's "comprehensive" scale; instead, they limit their analyses to single texts (a scale at which close reading remains putatively reasonable and uncontroversially "appropriate") and to relatively simple statistical calculations of surface features, using metrics that make no use of features derived from large-scale text collections. Their reluctance to wade into this variety of "distant reading" is perhaps explained by the disappointing results of much scholarship that adopts this perspective. Jockers's work in

Macroanalysis frequently fails to provide genuine critical insights; most often, he merely shows that his tools are working, employing them to support long-held critical commonplaces. The method Jockers describes in *Macroanalysis* for reliably predicting an author's gender based on thematic topic modeling data—female authorship, he finds, is predicted by engagement with stereotypically female themes such as "fashion" and "children"—does little, on its own, to enrich our understanding of literary history.[7] More disquieting still are the conclusions of a paper by Ryan Heuser and Long Le-Khac, former colleagues of Jockers in the Stanford Literary Lab. In their "Quantitative History of 2,958 Nineteenth-Century British Novels," they describe their discovery of two groups of words that exhibit exactly opposite frequency trends across their corpus: "abstract value" words such as "conduct" and "envy," which decrease steadily in the period; and "hard seed" words including action verbs, body parts, colors and numbers, which collectively increase in precisely inverse relation. The authors explain these shifts in terms of a turn in narrative style from telling to showing—a movement toward literary realism predicated on empirical description.[8] Yet research by Ted Underwood and Jordan Sellars offers a rather more mundane explanation[9]: they argue that *all* literary language experienced a shift from Latinate to Anglo-Saxon diction in the nineteenth century, and, noting that "abstract value" words tend to be Latinate and "hard seed" words Anglo-Saxon, suggest that the trend observed by Heuser and Le-Khac is indicative not of a shift from telling to showing, but simply evidence of a much broader change in literary fashion. Underwood and Sellars also note that the eighteenth century exhibits a reverse trend from Anglo-Saxon to Latinate diction. Since it would difficult to argue that this was the result of a shift in eighteenth-century narrative from showing to telling, Heuser and Le-Khac's explanation falls apart. Their paper thus stands as a manifest example of the dangers of working at the scale of big data, where, without close readings to ground interpretation, it is all too easy to impose grand theories on ambiguous results.

In *Macroanalysis,* Jockers, perhaps sensing that he has gone a step too far in calling close reading "totally inappropriate as a method of studying literary history," appends a conciliatory footnote. Citing Erich Auerbach's *Mimesis* (1948), he clarifies that he hasn't meant to "imply that scholars have been wholly unsuccessful in employing close reading to the study of literary history."[10] His mention of Auerbach at this point is significant, since Auerbach himself was extremely explicit about what he felt was the proper relationship between close and distant reading, big and small data,

in literary scholarship. "Philology and Weltlitatur" (1952), for example, reads as preemptive pre-digital rebuttal of distant reading techniques of the kind employed by Jockers, Heuser and Le-Khac. Beginning from "a great mass of material," Auerbach warns, will inevitably lead to "the introduction of hypostatized, abstract concepts of order."[11] At a scale where data is ambiguous and concrete close readings are in short supply, he argues, "ready-made, though rarely suitable, concepts whose appeal is deceptive because it is based on their attractive sound and their modishness, lie in wait, ready to spring in on the work of a scholar who has lost contact with the energy of the object of study."[12] For Auerbach, the solution lies in what we might now call a combination of close and distant reading.

Though Auerbach—the author of works with grandiose titles like *Mimesis: The Representation of Reality in Western Literature*—is very much interested in the "big," he argues that large-scale analysis must begin from concrete phenomena perceived in close reading. "In order to accomplish a major work of synthesis," he writes,

> it is imperative to locate a point of departure [*Ansatzpunkt*], a handle, as it were, by which the subject can be seized. The point of departure must be the election of a firmly circumscribed, easily comprehensible set of phenomena whose interpretation is a radiation out from them and which orders and interprets a greater region than they themselves occupy.[13]

In this chapter, we describe two projects that attempt an Auerbachian resolution of the close-versus-distant dilemma. Our "handle," our *Ansatzpunkt*, is modernist dialogism: the ethically charged, politically inflected tendency of modernist writers to include mutually differentially and often ideologically opposed voices in their works. Using cutting-edge techniques in computational stylistics, our work leverages the insights available at the scale of big data to model and explore dialogism as a concrete phenomenon in modernist texts. By developing new quantitative metrics that are trained on large datasets yet easily interpretable by humans, we build an important bridge between the scales of big and small data, and also between the disciplines of computer science and literary studies. Our approach is specifically tailored, moreover, to *modernist* literary studies, developing its computational style-based methodology in response to modernist-era accounts of the politics and ethics of genre (Mikhail Bakhtin's "dialogism" and Auerbach's "multipersonal representation of consciousness"). In our project *He Do the Police in Different Voices*,

which draws primarily on Bakhtin's account of dialogism, we use extrinsic features based on information from massive corpora to identify possible points of stylistic transition in T. S. Eliot's *The Waste Land*, and we employ our novel "stylistic profile" method to produce human-interpretable analyses of individual "voices" in the poem. In our project *The Brown Stocking*, which takes its theoretical impetus and its name from Auerbach's account of modernist polyvocality in the final chapter of *Mimesis*, we use stylistic profiles to analyze free indirect discourse (FID) and character speech in Virginia Woolf's *To the Lighthouse* and James Joyce's "The Dead." Our goal in these projects is not to produce definitive, computationally guaranteed readings, but rather to use computational analysis to test, probe and enliven human close readings. Rather than using distant reading to confirm broad critical metanarratives, we seek to establish a feedback loop in which the insights available at the scale of big data are employed to continuously challenge particular close readings. Ours is a hybrid approach that places distant and close reading in a reciprocal dialogue, based on the conviction that each stands to benefit from the perspective that the other has to offer.

Bakhtin, Auerbach and the Politics of Modernist Dialogism

The modernist period in Europe (roughly 1880–1950) was one of intense debate about the politics and ethics of genre, and the narratological theory of Mikhail Bakhtin and Erich Auerbach is representative of the modernist tendency to approach such questions through the lens of voice.[14] In "Discourse in the Novel," written in exile in Kazakhstan in the 1930s, Bakhtin championed the novel on the grounds that its multi-voiced and open-ended form presented a model of a pluralist, democratic society at a time of brutal totalitarian repression in the USSR.[15] Bakhtin's argument in favor of the novel is supported by an argument against poetry—particularly lyric poetry, which he positions as the novel's single-voiced other. Bakhtin's ideal novelist is one who renders the differentiated dialects of everyday life without seeking to order or purify them. In his account:

> The prose writer does not purge words of intentions that are alien to him, he does not destroy the seeds of heteroglossia embedded in words, he does not eliminate those language characteristics and mannerisms glimmering behind the words and forms.[16]

The poet, by contrast, purges, destroys, and eliminates in order to fit the linguistic universe into a single, standardized pattern. "The language of the poet is *his* language," Bakhtin writes: "he is utterly immersed in it, inseparable from it."[17] His primary formal example of what he calls the "single-personed hegemony"[18] of the poet's language comes in his analysis of rhythm, which, he argues, "destroys in embryo the social worlds of speech and of persons that are potentially embedded in word […] stripping all aspects of language of the accents and intentions of other people, destroying all traces of social heteroglossia and diversity of language."[19] For Bakhtin, formal categories of genre such as rhythm surpass the bounds of the merely aesthetic by modeling politically inflected modes of thought. The stylistically uniform lyric modeled acquiescence to totalitarianism, whereas the dialogic novel modeled open-ended democratic debate.

Just as Bakhtin wrote "Discourse in the Novel" in exile from an authoritarian regime, Auerbach, a German Jew, wrote *Mimesis* in exile from Nazi Germany. Like Bakhtin, Auerbach pursues in *Mimesis* a political reading of multi-voicedness. For Auerbach, what is most significant in twentieth-century fiction is its development of a technique he calls the "multipersonal representation of consciousness"[20]—a close analogue of Bakhtin's "dialogism." At a time when totalitarian regimes in Europe were violently imposing their single-voiced interpretations, Auerbach perceived a form of artistic resistance in modernist texts that offered:

> not one order and one interpretation, but many, which may either be those of different persons or of the same person at different times; so that overlapping, complementing, and contradiction yield something we might call a synthesized cosmic view.[21]

Auerbach's chief example of this bottom-up, multi-perspectival, multi-voiced conception of reality is Virginia Woolf's novel *To the Lighthouse*, in which, he argues, "the writer as narrator of objective facts has almost completely vanished" and "almost everything stated appears by way of reflection in the consciousnesses of the dramatis personae."[22] In Auerbach's reading of Woolf, one of the principal technical devices by which she achieves the "multipersonal representation of consciousness" is free indirect discourse (FID)—a narratological device for introducing character speech in such a way as to blur the boundaries between the voice of the narrator and that of the character, and so further to diminish narrator's role as a dispenser of authoritative truth. Together, Auerbach and Bakhtin present a powerful

case for dialogism as a crucial feature of modernist literature: a stylistic device, practiced and theorized by modernists themselves, seen to have social and ethical reverberations well beyond the sphere of the literary. Dialogism also provides an excellent starting point for reading modernism with machines: since style has proven historically to be the most tractable literary element for computational analysis, dialogic style presents a practical "handle" with which to grasp modernism digitally. Whereas many stylistic categories draw critics away from political or social contexts, dialogism draws us into a confrontation with the politics of form.

Modeling Dialogism in *The Waste Land:* Identifying Voice Switches

For reasons of space, time and language, T. S. Eliot had no access to the writings of either Bakhtin of Auerbach, yet he shared many of their concerns. Perhaps the most prominent lyric poet of the modernist period, he was sometimes attacked by his contemporaries for his inability or unwillingness to admit mutually differentiated, competing voices into his work; Virginia Woolf, for one, called him "a monologist."[23] In this respect, however, Eliot was perhaps his own harshest critic. Throughout the 1920s, Eliot repeatedly expressed his desire to abandon lyric poetry for a form even more multi-voiced than the novel: the narrator-less drama.[24] Eliot theorized the potential of dramatic form in essays of the period, such as "The Possibility of a Poetic Drama" (1920) and "Marie Lloyd" (1922), and experimented with it creatively in the hybrid poetic jazz drama provisionally titled *Wanna Go Home, Baby*, later published in fragmentary form as *Sweeney Agonistes* (1926–27). His best-known work, *The Waste Land* (1922), also bears the traces of Eliot's experiments with dramatic form, standing as a hybrid of his earlier lyric forms and the multi-voiced verse drama he would adopt from the 1930s onward. *The Waste Land* teems with voices—voices young and old, rich and poor, mundane and eternal, speaking all manner of languages and class dialects. Though Eliot does not provide a dramatis personae or mark the points of transition between the poem's voices, they emerge clearly in any good reading of the poem—for instance, in those by Alex Guinness and Fiona Shaw included on Faber's 2011 *Waste Land* iPad app. Though Eliot's own readings—two of which are included on the same app—are not nearly so vocally diverse, he clearly intended the poem to be understood as polyvocal, referring to

the "personage[s] in the poem" in his famous endnotes,[25] and employing the working title "He Do the Police in Different Voices."[26]

The latter is the name we have taken for our long-term project to explore and highlight the dialogism of this most famous of modernist poems.[27] Our work began in 2011 with the creation of a digital edition designed to emphasize the poem's uncertain generic status between single-voiced lyric and impersonal drama. The first stage involved aggregating 140 student interpretations of *The Waste Land* into a "class-sourced" reading of voices in the poem. Having asked students in "The Digital Text," a second-year English course at the University of Toronto, to indicate every instance in *The Waste Land* where they perceived a "voice switch," we used this data to devise a reading of the poem in which we identified sixty-eight voice switches and twelve characters.[28] On our project website, *hedothepolice.org*, we present this interpretation in the form of a digital edition ("What the Class Said" [WTCS]) that renders each unit of character speech in a unique typeface. The goal of this stage of the project was to teach students about modernist dialogism by having them *act it out*: to suggest that literary interpretation, particularly of dialogic modernist literary texts, is a communal, participatory act involving multiple competing perspectives. Crowdsourcing was thus employed not merely as a means to an end but also, to some extent, as an end in itself.

Taken on its own and isolated from the polyvocal process of its creation, however, the WTCS edition runs the risk of suggesting that its interpretation is definitive or "final." To mitigate against this suggestion, our project website includes an interactive page on which users can indicate their own set of voice switches and assign these to particular characters ("Have Your Own Say"). To further unsettle the particular interpretation of the WTCS edition, and to encourage further exploration of voices in *The Waste Land* generally, we have also sought to insert a computational "voice" into the discussion ("What the Computer Said"). Our work has pursued quantitative methods for performing the two basic interpretive tasks described to this point: first, segmenting the poem by identifying points where "voice switches" occur; second, clustering these discrete chunks into individual speaking voices.

Our first task was to develop a computational means of identifying the points in *The Waste Land* where one voice gives way to another. Our approach uses unsupervised techniques (that is, techniques that do not require human intervention at each step) in computational stylistics to locate instances of maximum "stylistic variation," using a procedure—described in further technical detail elsewhere[29]—that functions roughly

as follows. For every word in *The Waste Land*, we calculate a measure of stylistic change that takes into account a number of features in the spans of text immediately preceding and following that word. The features we consider fall into two categories: surface and extrinsic. Surface features, which are by far the more common for conventional computational analysis of literature (for instance, the work of Clement and Ramsay described above), can be calculated entirely from the text itself, requiring no external resources. These features include word length, syllable count, punctuation frequency, parts of speech, verb tense and type-token ratio (a measure of lexical density). Extrinsic features, which are more novel in analysis of literature, rely on lexical information derived from large external corpora. Such features include readability, sentiment polarity (the positive or negative affective stance of a given span), formality, and less human-interpretable (but extremely useful) features from latent semantic analysis (LSA). Our method works by investigating the features in a "sliding window" of text on either side of each word; for instance, it might calculate the sentiment polarity of the fifty words immediately preceding a given point in the text and compare it with the sentiment polarity of the fifty words following that point.[30] Our metric is built from the sum of the changes of all the features, and identifies voice switches at local maxima of the calculated change curve, such as the peaks represented on the curve in Fig. 3.1.

To test our method, we created artificial poems composed of randomly assembled sections of twelve poems of diverse style and authorship.[31] These artificial poems, with their unmarked transitions between styles and voices, mimic the stylistic diversity of *The Waste Land*. Our evaluation revealed that extrinsic features (particularly formality and LSA) slightly outperformed surface features in identifying transitions in our set of artificial poems, though the best results of all came from combinations of surface and extrinsic features. Next, we applied our method to two versions of *The Waste Land*: a "full" version containing all text in the poem except for headers and the dedication; and an "abridged" version omitting stanzas which are less than twenty words in length or in a language other than English—both conditions that make it difficult for our method to succeed.[32] Fig. 3.1 shows the change curve generated by our method for the abridged version of the poem, overlaying switches from the WTCS edition. In many instances, the switches identified by the algorithm coincide almost perfectly with those identified by human readers. Further, the model tends to predict more switches in sections where humans perceive numerous

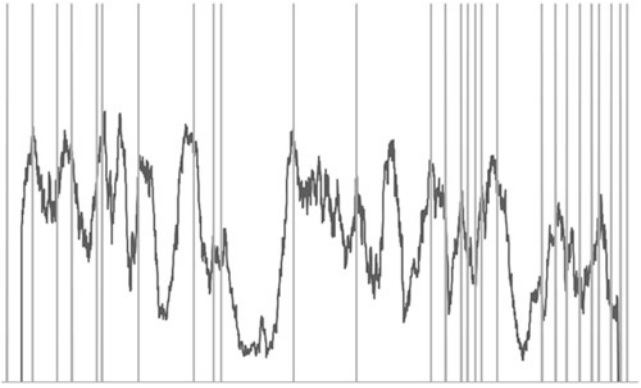

Fig. 3.1 Stylistic change curve over the abridged version of *The Waste Land*

switches, mostly notably in the last third of the poem. Our results thus bring us to the point at which computational analysis trained on large datasets can meaningfully begin to contribute to close reading: the results conform sufficiently to human interpretation to convince us that they are not merely random, yet diverge sufficiently to allow us to evaluate whether particular machine interpretations can offer something new.

On closer inspection, we found that several of the points at which the computational model departs from the human interpretation suggests new and insightful interpretations. Consider the famous opening lines of the poem:

> April is the cruellest month, breeding
> Lilacs out of the dead land, mixing
> Memory and desire, stirring
> Dull roots with spring rain.
> Winter kept us warm, covering
> Earth in forgetful snow, feeding
> A little life with dried tubers.
> Summer surprised us, coming over the Starnbergersee
> With a shower of rain; we stopped in the colonnade,
> And went on in sunlight, into the Hofgarten,
> And drank coffee, and talked for an hour. (1–11)

Here, our human reading placed a switch between lines 4 and 5: we attributed the first lines to the narrator figure we named Tiresias, and

the next lines to the aristocratic character we named Marie, largely on the assumption "us" of line 5 refers to the latter's family on vacation in Switzerland. The machine model, however, places the switch between lines 7 and 8. On reflection, it does so with good reason. While the passage from lines 5–7 transitions gradually between the dreary, remote tone of the opening lines ("cruellest," "dead land") and the more neutral, slightly hopeful tones of lines 8–11 ("shower of rain," "went on in sunlight"), the negative tone of the opening lines remains palpable in phrases like "forgetful snow" and "dried tubers." As Michael Levenson argues in *A Genealogy of Modernism*, "the stylistic patterns shifts" between lines 7 and 8: in a human close reading that relies heavily on computationally tractable "surface features" such as syntax and verb tense, Levenson notes, "The series of participles disappears, replaced by a series of verbs in conjunction" and "The adjective-noun pattern is broken."[33] Reconsidering the passage, we agree with the close readings of Levenson and of the algorithm that we ought to have placed the switch between lines 7 and 8.

The computer model suggests another insightful interpretation in the following stanza:

> Unreal City
> Under the brown fog of a winter noon
> Mr. Eugenides, the Smyrna merchant
> Unshaven, with a pocket full of currants
> C.i.f. London: documents at sight,
> Asked me in demotic French
> To luncheon at the Cannon Street Hotel
> Followed by a weekend at the Metropole. (207–14)

Our interpretation attributed the entirety of this stanza to the character we (the first author and the 140 students) named "Crazy Prufrock," the educated but increasingly unhinged character who earlier in the poem speaks of planting a corpse in a garden (60–76). We attributed this stanza to him largely because it begins with the words "Unreal City," the same phrase that opens the account of the corpse. While we allowed this opening phrase to color our interpretation of the rest of the stanza, the computer model inserts a break early in the paragraph, between lines 208 and 209. On reflection, this seems to us a preferable interpretation, since while the first two lines have a Prufrockian air, the remainder of the stanza is delivered in a balanced, detached tone more reminiscent of the poem's narrator

figure, Tiresias. Notably, Jewel Spears Brooker and Joseph Bentley also attribute this passage to Tiresias, following an exhaustive and ingenious close reading, in which they determine him to be "the only figure in the vicinity of the poem who can be trusted to see all about the figure [of Mr. Eugenides]."[34] In instances like these, our model shows the power of computational methods trained on massive corpora to contribute productively to the minutest of human close readings.

Modeling Dialogism in *The Waste Land*: Stylistic Profiles

Having segmented the poem into stylistically distinct chunks, our second computational task for *The Waste Land* involved clustering: determining whether we could group passages belonging to distinct characters. Pursuing a similar LSA-centered feature set to that we used for segmenting the poem, our work was moderately successful.[35] Since LSA produces results that are not particularly interpretable by humans, however, this work did little to further our goal of provoking new close readings. Realizing that a new metric was required in order to produce the sorts of interpretations that might prompt new close readings, we turned our focus toward developing our signature six-dimensional approach to linguistic style, which employs an automatically created lexical resource to produce human-interpretable "stylistic profiles."[36] As our work proceeded in *The Waste Land*, we came to realize that we could use these stylistic profiles as a means of testing our intuitions in assigning particular spans of text to particular characters. The advantage of this approach—the reason we have found it so useful for literary analysis and the reason we believe it represents a large step forward from techniques like PCA and LSA—is its accessibility and transparency even to readers entirely unversed in computational stylistics.

Our profiling method is based on six discrete aspects of style: objectivity (use of words that project a sense of disinterested authority, such as *invariable* and *ancillary*); abstractness (words denoting concepts that cannot be described in purely physical terms, and which require significant cultural knowledge to understand, such as *solipsism* and *alienation*); literariness (words found in traditionally literary texts such as *wanton* or *yonder*); colloquialness (words used in informal contexts such as *booze* and *crap*); concreteness (words referring to events, objects or properties in the physical word, such as *radish* and *freeze*); and subjectivity (words that are

strongly personal or reflect a personal opinion, such as *ugly* and *bastard*). Our process for building stylistic lexicons, described in detail elsewhere,[37] functioned as follows. First, we produced a list of approximately 900 words carefully selected for their stylistic diversity, which human annotators (five university-educated native English speakers trained for the task with simple written guidelines) evaluated in terms of the six stylistic aspects listed above. (Annotators noted, for instance, that the word "brazen" projected subjectivity and literariness but none of the other aspects.) Once sufficient inter-annotator agreement was reached, we used an automated procedure to collect information on how these 900 words are employed in a large corpus composed of all English texts in the 2010 image of *Project Gutenberg*.[38] Using this information, we were able to derive stylistic information automatically for *any* word; in this case, we investigated every word in *The Waste Land* (as well as significant multi-word expressions, such as *from time to time* and *ought to be ashamed*[39]) and, based on their employment in the same *Project Gutenberg* corpus, assigned a value between −1 and +1 (to twelve decimal places) for each of the six stylistic aspects. Using this information, we are able to produce stylistic profiles for particular segments and particular characters in the poem by aggregating results for individual words or multi-word expressions.

This method proved extremely successful in capturing individual characters' manners of speech in the WTCS reading of the poem (Table 3.1).

Human readers often identify Woman in Bar, the Cockney woman whose speech dominates the end of "A Game of Chess," as the most distinctive voice in the poem. Our computational approach likewise found her voice to be the most distinct. Her stylistic profile—marked by extremely high colloquial and subjective values, and extremely low values for the objective and literary dimensions, all of which corresponds to our intuitions—is distinguishable from all other voices in the poem, in most of the six aspects, at statistical significance of $p < 0.001$ (where $p < 0.05$ is considered a reliable threshold of statistical significance). The stylistic profiles of other characters likewise conformed to our qualitative expectations. Marie, emotional and nostalgic with highly oral language, is marked by high subjectivity and high colloquialness. Crazy Prufrock, educated but unbalanced, is marked in our analysis by high abstraction, high colloquialness and high objectivity, indicating not mental stability but high cultural knowledge and education. The narrator figure we call Tiresias is marked by relatively low values for colloquialness and correspondingly high values for objectivity and literariness.

Table 3.1 Stylistic profiles for various characters in *The Waste Land*

Character	*Unique words*	*Stylistic dimensions*					
		Objective	*Abstract*	*Literary*	*Colloquial*	*Concrete*	*Subjective*
Tiresias	460	0.09	−0.04	0.03	−0.21	0.02	−0.03
Marie	132	−0.07	−0.13	−0.03	0.02	0.04	0.03
Hellfire preacher	207	0.00	0.00	0.06	−0.14	0.07	−0.06
Chorus	105	−0.02	−0.02	0.03	0.14	0.04	−0.06
Intrepid reporter	66	−0.06	0.28	−0.03	0.07	−0.06	0.05
Madame sosostris	15	−0.47	0.26	−0.08	0.65	−0.14	0.16
Crazy prufrock	399	0.01	0.07	0.00	0.01	−0.01	−0.01
Nervous one	126	−0.26	0.09	0.04	0.29	−0.07	0.07
Woman in bar	151	−0.45	−0.01	−0.24	0.73	−0.11	0.20
The typist	54	0.14	0.42	−0.03	−0.14	−0.02	−0.02

These values show us that our method seems to work; that is, that it produces human-interpretable results that correspond sufficiently to our intuitions to enable us to trust them. Yet, as we argue of all computational metrics, stylistic profiles only really become useful when they suggest something we didn't already know. In this case, they proved useful by prompting us to reconsider our assignments of particular passages to specific characters. One nagging concern we encountered in devising our human interpretation of the poem was whether Tiresias and Crazy Prufrock were sufficiently distinguishable to stand as independent characters. Given that both voices were marked by the same qualitative traits—wordiness, a deep familiarity with the literary tradition and a fondness for literary quotation—we sometimes wondered, along with critics like Calvin Bedient,[40] whether they weren't simply projections of a single consciousness. Our stylistic profiles provide reason to consider the two characters distinct. Similar as the voices are in the literary dimension, they are strongly distinguished in colloquial ($p < 0.001$), where Prufrock's schizophrenic shifts across registers produce much higher values. We likewise debated whether Crazy Prufrock is speaking to himself or to another voice in the extended back-and-forth dialogue that occurs in the middle of "A Game of Chess." In our WTCS interpretation, we described this passage as an exchange between Prufrock and another character, "Nervous One," and data from our stylistic profiles reinforces our choice by strongly distinguishing the voices in the subjective, objective and colloquial dimensions (all $p < 0.001$).

Stylistic profiles were perhaps most useful of all for testing our qualitative "clustering" of the poem, certainly the most subjective and intuitive interpretive procedure we employed to produce the WTCS edition. In a few instances, data that seemed to suggest a misreading in our interpretation in fact reinforced it. Despite strongly divergent style data for the second (77–110) and third (215–56) passages we attributed to Tiresias ("Tiresias 2" and "Tiresias 3" in the naming convention followed in the rest of the chapter), for instance, we remain convinced of our reading. In Tiresias 2, which describes a rich woman's elaborate grooming ritual, the narrator's presentation is strongly ironic: the evocation of "The chair she sat in, like a burnished throne," borrowed from Enobarbus's account of Cleopatra's raft in *Antony and Cleopatra*, is deliberately overblown, serving to demonstrate the extreme disconnect between the cocoon of the dressing-room and the "Unreal City" beyond. This disjuncture is signaled through the painting that sits on the woman's mantel, depicting a scene from the Philomela

myth. Not even a pastoral rendering of the story can avoid evoking the brutality of Tereus's rape, and in presenting his *ekphrasis*, Tiresias momentarily abandons his hyper-refined diction to comment with unadorned lexis on the persistence of cruelty in the modern world: "And still she cried, and still the world pursues,/'Jug Jug' to dirty ears" (102–3). By contrast, Tiresias 3, in which the self-assured "young man carbuncular" forces himself upon the passive "typist," presents a much more direct account of a contemporary rape. While these scenes clearly respond to and mirror one another, they yield very different stylistic profiles. The language in Tiresias 3 is significantly more colloquial than in Tiresias 2 ($p < 0.01$), reflecting the flatly sordid account of the typist's rape. It is also markedly more subjective ($p < 0.01$), reflecting the more honest account of the typist's feelings, as opposed to the ironic evocation of the rich woman's hermetic emotional landscape, buffered on all sides by luxury. In this case, then, the divergent stylistic profiles simply highlight the chameleonic aspect of Tiresias's narratorial style, which adapts itself to the particular scene presented. [41]

Elsewhere, however, style data led us to change our interpretation. Another discrepancy in passages attributed to Tiresias—between Tiresias 2 and Tiresias 5 (378–85)—uncovered an untenable reading. Tiresias 5 begins with a description that recalls the dressing scene ("A woman drew her long black hair out tight"); another link is established between the passages through the echo of the opening words of Tiresias 2, "At the violet hour," in the description of "bats with baby faces in the violet light" (380). Yet while such reverberations were sufficient to convince us of a connection, our stylistic profiles show little to suggest a common speaker. Prompted by this data, we reconsidered the passage, and noted that it echoes words not only from Tiresias, but also from numerous other voices in the poem. Its evocation of "towers/Tolling reminiscent bells, that kept the hours" (384), for instance, recalls two passages we attributed to Crazy Prufrock: "where Mary Woolnoth kept the hours" (Prufrock 2, 67) and "Falling towers" (Prufrock 15, 374). Since this passage deliberately mixes together fragments of voices from throughout the poem, we decided to attribute this passage to the non-personal entity we call "The Chorus."

Beyond testing particular interpretations, stylistic profiles can provide a starting point for evaluating writers' representations of certain classes of characters. For instance, we were interested to see whether Eliot's male or female voices are more mutually differentiated. Investigating the figures, we noted that his male characters are more vocally diverse, and that each of his female characters (Marie, Madame Sosostris, Nervous One,

Woman in Bar and The Typist) has a relatively high score for subjectivity, possibly indicating a stereotyped representation.[42] Certain female voices are quite distinct: Marie and Madame Sosostris, for example, register statistically significant differences in abstract, colloquial and concrete (all $p < 0.01$). Yet, in a poem that differentiates so successfully between its voices—across all possible pairings of characters, only two pairings fail to register a single statistically significant difference ($p < 0.05$)—it is telling that one of these indistinct pairings should be between female characters, Nervous One and Madame Sosostris. (The other is Crazy Prufrock and the non-personal Chorus.) Yet careful analysis is required before we jump to conclusions: their similarity may be due to Eliot's failure to distinguish female voices, but it may also be due to these characters' similar registers (both highly oral) or simply to the fact that there is insufficient data for Madame Sosostris, who speaks very little. Although the stylistic profiles we produced for *The Waste Land* were not able to answer these questions definitively, they were able to raise them with new urgency. As such, they were sufficiently promising to prompt us to investigate their application in other modernist texts.

Quantifying Free Indirect Discourse in *To the Lighthouse* and "The Dead"

At the time of our investigation of *The Waste Land*, we were involved in another project focused on modernist dialogism, *The Brown Stocking*, which looked at free indirect discourse (FID) in Virginia Woolf's *To the Lighthouse* and James Joyce's "The Dead." (The name of the project is taken from the final chapter of Auerbach's *Mimesis*, where he reads FID in *To the Lighthouse* as an example of modernist "multipersonal representation of consciousness.") Though this project was not initially devised with stylistic profiles in mind, it benefitted significantly from a shift toward a style-based approach, further demonstrating the power of computational methods trained on large-scale datasets to vivify literary inquiry and contribute meaningfully to close reading.

We began *The Brown Stocking* with three principal aims. First, we wanted to help our undergraduate students better understand *To the Lighthouse* by highlighting its principal interpretive dilemma: the vexed question of *who is speaking* at any given point. We pursued this through a TEI encoding exercise that asked students to annotate short passages from the novel. For each instance of character speech in their assigned passage, students were

asked to indicate whether it was introduced as direct, indirect or free indirect discourse; whether it was spoken aloud or silently; and which character was speaking. Because there are often multiple valid interpretations of a given passage, we assigned each to four or five students. We devised this as an exercise in computer-assisted close reading, and, in practice, students reported that the act of translating their implicit interpretations into explicit markup helped them to clarify their reading of the text. The next goal was to combine these interpretations into a digital edition of *To the Lighthouse* that would serve as a "reader's map," showing the vast array of possible interpretations of Woolf's text and thus visualizing an active circuit of modernist dialogism: a dialogic novel prompting a dialogic scene of reader response. Following two rounds of annotation, each involving approximately 160 students and focusing respectively on the first four and final seven chapters of the novel, we published this edition on the project website, *brownstocking.org*.[43]

Our final goal was to devise a means of using these student annotations to train a machine-learning model that could detect FID automatically in untagged plain text. In pursuing this goal, we were consciously seeking to replicate Auerbach's understanding of the "multipersonal representation of consciousness" as an aggregation of numerous distinct interpretations that, when combined, provide a "synthesized cosmic view." In practice, however, this provided difficult because inter-annotator agreement was quite low, due to the complicated, multi-voiced nature of the text, in which Woolf uses FID so pervasively. We thus decided to perform another round of annotation on a modernist text with a more conventional use of FID—James Joyce's "The Dead"—yet the added data brought us no closer to a machine-learning system for detecting FID. (We were, however, able to devise a relatively accurate rule-based system for identifying FID from grammatical and syntactic clues—and we produced a "reader's map" edition for "The Dead" at *livingdead.ca*).[44] The data proved immensely useful, however, in a task quite different from that for which it was initially collected: the further exploration of our method of stylistic profiling.

The first research question we posed was a fundamental one related to the definition of FID. If FID has become today a reasonably familiar element of literary discourse, the history of the invention, detection, or critical elaboration of FID is sufficiently curious to merit careful scrutiny. If we consider that the first novels were produced in the sixteenth century, it took some two hundred years of literary history for FID to first be employed; though Cervantes used direct and indirect discourse, it was

not until the time of Austen and Goethe that FID appeared in the novel. Following its invention, it took another century for critics to notice it. It is generally agreed that Adolph Tobler was the first to identify the device, calling it "a peculiar mixture of direct and indirect speech" in 1892.[45] In the years that followed, FID became a focus of intense modernist critical scrutiny. Graham Pechey estimates no fewer than eighteen separate names were given to the device in the modernist period, among them "veiled speech" (Theodor Kalepky 1912), "free indirect style" (Charles Bally 1912), "pseudo-objective speech" (Leo Spitzer 1921) and "pseudo-objective discourse" (Mikhail Bakhtin, 1920s).[46] The one common notion in these various definitions of FID—a notion that retains its critical force today—is that FID is an "in-between" mode of discourse (a "peculiar *combination*," a "*pseudo*" or "*veiled*" form) existing on the continuum between pure narration and direct discourse. Given the delayed and uncertain process of defining FID—a process that was carried out in a haphazard and entirely qualitative manner—we were interested to see whether our quantitative method could support or refute the notion of FID's "in-betweenness."

Our method for testing this definition, described in further technical detail elsewhere,[47] proceeded as follows. First, we located all the passages in *To the Lighthouse* and "The Dead" in which a majority of annotators identified a span as FID, direct discourse, or narration.[48] Then, using the method described earlier in relation to *The Waste Land*, we built stylistic lexicons for both texts, and used these to generate stylistic profiles for narration, spoken direct discourse, silent (thought) direct discourse, and FID (Table 3.2).

For both texts, our results largely conformed to expectations. In "The Dead," FID is "in-between" in all six dimensions, most clearly in colloquial and subjective. In *To the Lighthouse*, FID occupies a middle position in four of six stylistic dimensions: it is more abstract than narration, but less abstract than directly rendered thought; more literary than narration but less so than direct speech or thought; less concrete than narration but more so than direct speech or thought; and so on. Exceptions occur in objective and colloquial, where FID is in an extreme position; yet in both cases, FID tracks closely with narration, and the particular divergences may simply reflect a mannerism of Woolf's narrator, who tends not to admit colloquialisms when mixing her language with that of her characters. Our work thus offers quantitative support for two long-held but seldom-tested hypothesis about FID: that it is an identifiable mode of discourse distinct from narration and direct discourse, and that it falls stylistically between these two poles.[49]

Table 3.2 Stylistic profiles for discourse types in *To the Lighthouse* and "The Dead"

Text	*Discourse*	*Unique words*	*Stylistic dimensions*					
			Objective	*Abstract*	*Literary*	*Colloquial*	*Concrete*	*Subjective*
To the Lighthouse	Narrator	765	0.00	0.00	0.00	0.00	0.00	0.00
	FID	2916	0.08	0.16	0.02	−0.02	−0.15	0.02
	Thought	212	−0.15	0.21	0.07	0.30	−0.20	0.08
	Speech	172	−0.32	0.14	0.06	0.49	−0.20	0.11
"The Dead"	Narrator	1325	−0.01	0.02	0.08	0.04	0.00	0.09
	FID	400	−0.13	0.19	0.10	0.19	−0.15	0.11
	Thought	57	−0.43	0.18	0.12	0.74	−0.30	0.22
	Speech	651	−0.11	0.23	0.06	0.27	−0.19	0.16

Next, we investigated whether stylistic profiles would prove insightful in the mixed, murky waters of FID. Although the method worked well in *The Waste Land*, where speech is rendered mostly as direct discourse, we were unsure whether individual characters' stylistic personalities would reveal themselves in FID, in which their speech is mixed with that of the narrator. Here, again, the results were promising. The stylistic profiles of the narrators of *To the Lighthouse* and "The Dead" in Table 3.2 highlight revealing differences: where Woolf's narrator is consistently flat, detached and objective, Joyce's narrator scores higher values for literary, colloquial, and subjective.

As Table 3.3 shows, stylistic profiles also provide insights into the FID of individual characters. Gabriel's distinct manner—reserved, given to deep thoughts and literary quotation—emerges clearly in the profile of his FID, which is notably less colloquial, more literary and more abstract than that of other characters. Profiles of Woolf's FID are likewise revealing. Some of the rift between Mr. and Mrs. Ramsay is captured in their style values: Mr. Ramsay, lost in his world of philosophical speculation and scholarly research, is much more abstract, much more literary and far less concrete than Mrs. Ramsay. Most interesting in *To the Lighthouse* are the relationships of inter-character influence that the stylistic profiles suggest. Though their style profiles are quite dissimilar, Mr. and Mrs. Ramsay have much more in common with one another stylistically than they do with their children. Cam and James spend much of Part III of the novel pondering the influence of their parents, mourning the lost influence of their mother while bristling against the domineering authority of their father. This generational conflict is expressed at the level of style: the Ramsays and their children speak very different languages, with the latter notably less objective, less abstract and more colloquial. Lily too ponders the influence of the elder Ramsays in Part III. She is particularly ambivalent toward Mrs. Ramsay, whom she admires deeply while resisting the conventional gender role she adopts in her family. Despite these misgivings, stylistic profiles suggest Lily is indeed Mrs. Ramsay's stylistic heir: in all six dimensions, their profiles are nearly identical. The situation is very different for Mr. Ramsay and his would-be protégé, Charles Tansley. The young philosopher is desperate to belong to the Ramsays' social and intellectual world, yet is bitterly aware of the barrier that his working-class origins present. Tansley's failure to integrate himself into their sphere is as plain in the stylistic profile data as in the plot of the novel. Stylistically, Tansley and Mr. Ramsay are extremely dissimilar with the exception of their shared philosophical penchant for the abstract.

Table 3.3 Stylistic profiles for characters in *To the Lighthouse* and "The Dead"

Text	*Character*	*Unique words*	*Stylistic dimensions*					
			Objective	*Abstract*	*Literary*	*Colloquial*	*Concrete*	*Subjective*
To the Lighthouse	Mrs. Ramsay	805	0.07	0.24	0.00	0.03	−0.22	0.03
	Mr. Ramsay	70	0.09	0.58	0.27	0.01	−0.49	0.00
	William Bankes	248	−0.01	0.19	0.03	0.14	−0.17	0.08
	Lily Briscoe	1485	0.06	0.17	0.03	−0.02	−0.15	0.01
	James Ramsay	540	−0.06	0.03	0.08	0.06	−0.03	0.02
	Cam Ramsay	381	−0.10	0.04	0.04	0.10	−0.06	0.00
	Charles Tansley	138	−0.07	0.21	−0.07	0.22	−0.23	0.05
"The Dead"	Gabriel	358	−0.12	0.21	0.12	0.17	−0.17	0.10
	Other	85	−0.30	0.06	0.02	0.44	−0.07	0.14

Table 3.4 Stylistic profiles for various social groups in *To the Lighthouse*

Social identity		*Unique words*	*Styles*					
Factor	*Category*		*Objective*	*Abstract*	*Literary*	*Colloquial*	*Concrete*	*Subjective*
Age	Young	969	**−0.03**	**0.06**	0.04	**0.04**	**−0.06**	0.01
	Old	2248	**0.09**	**0.21**	0.02	**−0.02**	**−0.19**	0.02
Class	Lower	138	**−0.07**	0.21	**−0.07**	**0.22**	−0.23	0.05
	Higher	2844	**0.08**	0.16	**0.03**	**−0.02**	−0.15	0.02
Gender	Female	2356	**0.08**	0.18	0.02	−0.02	−0.17	0.01
	Male	878	**0.02**	0.14	0.06	0.03	−0.12	0.03

Note: Bold indicates statistically significant difference at the $p < 0.01$ level between two categories of the same factor.

Prompted by the disparity between Tansley and Mr. Ramsay, we pursued a deeper investigation into the influence of socioeconomic categories on character speech in *To the Lighthouse* (Table 3.4). Tansley is the only character of working-class background who is attributed FID in the novel; working-class characters such as Macalister or Mrs. Beckwith speak directly or not at all. Comparing his limited FID with that of characters of higher class, however, we find a conventional power dynamic at work: higher-class characters are more authoritative, more literary, more concrete, less subjective and far less colloquial. The stylistic difference between age groups is similar to that between classes, though the key distinction for age is abstraction (words that require significant cultural knowledge) whereas that for class is literariness. Perhaps most interesting is that while *To the Lighthouse* reproduces conventional power dynamics for class and age, it almost completely reverses them for gender. Compared to male characters, female characters are more objective, more abstract, less colloquial and less subjective—and Mr. Ramsay's extreme values for literariness and concreteness likely explain why men rank slightly above women in these categories.

These results may be taken by some critics as confirmation of biases in Woolf's authorial practice. As one of the most vigorous champions of feminism and female authorship of the modernist period, it will come as little surprise that she extended this struggle to the level of style, erasing and indeed reversing gendered linguistic power dynamics. On the other hand, those who have accused the upper-middle-class Woolf of insensitive or stereotyped representations of lower-class characters[50] will find quantitative support in our stylistic profiles. As elsewhere, however, we urge readers to consider these figures not as the final word, or definitive proof,

but rather as prompts for further close reading. Indeed, Bakhtin and Auerbach championed modernist dialogism precisely because, by presenting markedly differentiated strata of socially inflected speech, it modeled the lively interchange of democratic debate. From our perspective, the question of whether Woolf should be applauded or condemned for differentiating the speech of characters of different ages, classes and genders is one that requires more than quantitative data to answer. How mimetically accurate is her depiction of female or lower-class speech? What resources, qualitative and quantitative, might we need to draw upon in determining this? Where she departs from mimesis, how likely is it that she does so deliberately? If her departure is deliberate, what is she trying to achieve? If not, how might this lead us to re-evaluate her authorial practice, or modernist authorship more generally? Responding to these questions, raised by computational models trained on large data sets, requires all our resources as literary critics: intimate familiarity with literary history, knowledge of context, and the ability to read closely and carefully.

Big Data in the Hermeneutic Circle

In this chapter, we've focused on the way that analytic techniques trained on large datasets can animate interpretation of a few canonical modernist texts; as our subtitle suggests, we have looked at "close reading with big data." As we reflect on what we've learned in our research, our focus is shifting toward applying these techniques to ever-larger numbers of texts. In order to build the stylistic lexicons we used to produce stylistic profiles of voices in *The Waste Land*, *To the Lighthouse* and "The Dead," we developed a technique for automatically separating character speech from narration in untagged plain text.[51] Applying our rule-based approach to identifying FID, and supplementing it with what we learned from investigating stylistic profiles in Eliot, Woolf and Joyce, we are now developing techniques for automatically identifying characters and classifying their speech as direct, indirect and free indirect discourse. Having demonstrated the usefulness of our method of stylistic profiles through close engagement with individual literary texts, we are in a position to begin an algorithmic investigation of the history of dialogism in English-language fiction. Now that we are able to derive automatic dramatis personae for any novel or play, and to calculate a quantitative measure of the stylistic diversity that exists in each text, we will have a quantitative means gaining insight into several large-scale questions about dialogism. Are the works of modernist writers like Woolf, Joyce and Eliot—all of whom pursued dialogism as a conscious aim—really

the most dialogic in the literary record? How does dialogism map onto historical time; do periods of political turmoil correspond to changes in the stylistic diversity of fiction? Which regions produce the most stylistically varied writing? Do changes in the dialogism of fiction anticipate changes in non-fiction? What previously ignored authors, periods and genres might our method consider as particularly dialogic? As we make this Auerbachian leap from the concrete "handle" of modernist dialogism to the largest scale of literary history, we expect our technique to raise new questions, to prompt investigations of new texts, and to alert us to unexpected writers, periods and genres—in other words, to supply us with an abundance of material that will require our most attentive close reading.

In pursuing our research, we find it useful to envision the role of computational analysis within the framework of Dilthey's hermeneutic circle. Dilthey posits that literary interpretations emerge from interactions at different scales of meaning: the movement of the hermeneutic circle is propelled by the paradoxical fact that while we can understand the whole of a literary work only through careful consideration of its individual parts, so too can we know individual parts only through careful consideration of whole. In the hermeneutic circle, literary interpretation is a necessarily mobile, dynamic act of holding together various mutually interdependent elements. From our perspective, the insights available at the scale of big data contribute to, and by no means invalidate, this dynamic. To shift metaphors somewhat, we see big data as a cog in the movement of the hermeneutic circle rather than a wrench thrown into the works. In our investigations into modernist dialogism, extrinsic features and human-interpretable stylistic profiles trained on massive datasets helped us to refine our interpretations, shed light on fine points of theme and characterization, and allowed us to probe basic definitions of literary terms. In each of these tasks, close and distant reading are complementary. Far from "inappropriate" in the context of big data, close reading remains the ground by which distant reading achieves its effects and demonstrates its usefulness.

Notes

1. Matthew L. Jockers, *Macroanalysis: Digital Methods and Literary History* (Urbana: University of Illinois Press, 2013), 7.
2. Ibid.
3. Susan Hockey, "The History of Humanities Computing," in *A Companion to Digital Humanities*, ed. Susan Schreibman, Ray Siemens and John Unsworth (Oxford: Blackwell, 2001), http://perma.cc/L3A8-EJHV.

4. Julia Flanders, "Detailism, Digital Texts, and the Problem of Pedantry," *TEXT Technology* 14, no. 2 (2005): 57.
5. Stephen Ramsay, *Reading Machines: Toward an Algorithmic Criticism* (Urbana, IL: University of Illinois Press, 2011).
6. Tanya Clement, "Text Analysis, Data Mining, and Visualizations in Literary Scholarship," in *Literary Studies in the Digital Age: An Evolving Anthology*, ed. Kenneth M. Price and Ray Siemens (MLA Commons, 2013), http://perma.cc/2CED-BNEK.
7. Jockers, *Macroanalysis*, 151–3. Jockers's more recent work employing sentiment analysis to identify six basic plot shapes has been variously attacked as flawed, misguided and reductive, as well as praised as path breaking and insightful. Until Jockers publishes his results and the scholarly community has an opportunity to test them, it remains too early to take a side. See Eileen Clancy, "A Fabula of Syuzhet: A Contretemps of Digital Humanities and Sentiment Analysis," *Storify*, May 3, 2015, https://storify.com/clancynewyork/contretemps-a-syuzhet.
8. Ryan Heuser and Long Le-Khac, *A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method*, Stanford Literary Lab Pamphlets 4 (Stanford, 2012), http://litlab.stanford.edu/LiteraryLabPamphlet4.pdf.
9. Ted Underwood and Jordan Sellers, "The Emergence of Literary Diction," *Journal of Digital Humanities* 1, no. 2 (Spring 2012), http://perma.cc/K655-GMLG.
10. Jockers, *Macroanalysis*, 7. Jockers later (p. 26) goes further, arguing that he is "not suggesting a wholesale shelving of close reading" but in fact recommending "a blended approach." Although he writes that close reading and distant reading are "not antithetical"—and indeed "share the same ultimate goal of informing our understanding of the literary record"—his work implies, at best, a model of close and distant reading as *parallel* interpretive strategies, whereas our argument is that they are most productive when positioned as a "feedback loop."
11. Erich Auerbach, "Philology and Weltliteratur," trans. Edward Said and Marie Said, *The Centennial Review* 13, no. 1 (Winter 1969): 10.
12. Ibid., 16.
13. Ibid., 13–14.
14. For more on Bakhtin, Auerbach, and a historicized account of the development of their theories of polyvocality, see Adam Hammond, "The Honest and Dishonest Critic: Style and Substance in Mikhail Bakhtin's 'Discourse in the Novel' and Erich Auerbach's *Mimesis*," *Style* 45.4 (Winter 2011): 638–53.
15. Ken Hirschkop, *Mikhail Bakhtin: An Aesthetic for Democracy* (Oxford ; New York: Oxford University Press, 1999).

16. Mikhail Bakhtin, "Discourse in the Novel," in *The Dialogic Imagination: Four Essays*, ed. Michael Holquist, trans. Caryl Emerson and Michael Holquist, 18. paperback printing, University of Texas Press Slavic Series 1 (Austin, Tex: Univ. of Texas Press, 2011), 298.
17. Ibid., 285.
18. Ibid., 297.
19. Ibid., 298.
20. Erich Auerbach, *Mimesis: The Representation of Reality in Western Literature*, trans. Willard Trask (Garden City, NJ: Doubleday, 1957), 536.
21. Ibid., 549.
22. Ibid., 534.
23. Virginia Woolf and Anne O. Bell, *The Diary of Virginia Woolf. Vol. 5: 1936–1941*, 1. Harvest ed. (New York, NY: Harcourt Brace & Company, 1985), 210.
24. For an influential contemporary genre analysis that privileges drama for its multi-voicedness, see Stephen Dedalus's famous discussion in James Joyce, *A Portrait of the Artist as a Young Man* (New York: B. W. Huebsch, 1921), 251–2.
25. T. S. Eliot, *The Waste Land*, in *Collected Poems, 1909–1962* (London: Faber and Faber, 1963), 82.
26. For more on the Eliot's working title, see Craig Raine's video essay in *The Waste Land* (London: TouchPress, Faber and Faber, 2011).
27. Adam Hammond and Julian Brooke, *"He Do the Police in Different Voices: Exploring Voices in T. S. Eliot's* The Waste Land," 2012, http://hedothepolice.org/.
28. Students provided their readings of the poem *before* being "taught" the poem in lecture. Most students enrolled in the class had previously studied the poem in their first-year introductory classes. While their readings were not colored by the lectures in "The Digital Text," we cannot discount the effect of earlier instruction.
29. Julian Brooke, Adam Hammond and Graeme Hirst, "Distinguishing Voices in *The Waste Land* Using Computational Stylistics," *Linguistic Issues in Language Technology* 12.2 (October 2015): 1–43.
30. The actual size of the "window" is given as w in our paper "Distinguishing Voices in *The Waste Land* Using Computational Stylistics."
31. The poems are W. H. Auden, "September 1, 1939"; Rupert Brooke, "Wagner"; T. S. Eliot, "The Love Song of J. Alfred Prufrock"; D. H. Lawrence, "Ballad of Another Ophelia"; Mina Loy, "Giovanni Franchi"; Wilfred Owen, "Strange Meeting"; William Shakespeare, "How Should I Your True Love Know?" (Ophelia's song from *Hamlet*); Stevie Smith, "Not Waving but Drowning"; Edmund Spenser, "Epithalamion"; Algernon Charles Swinburne, "Before the Beginning of Years"; Alfred,

Lord Tennyson, "The Coming of Arthur"; and Dylan Thomas, "A Saint About to Fall."

32. For a more detailed explanation, see Brooke, Hammond and Hirst, "Distinguishing Voices in *The Waste Land* Using Computational Stylistics."
33. Michael Levenson, *A Genealogy of Modernism*, (Cambridge: Cambridge University Press, 1984), 36. Though Levenson initially notes that a switch seems to occur between lines 7 and 8, and not between lines 4 and 5, his larger point is that the poem's "overlapping principles of similarity undermine the attempt to draw boundaries around distinct speaking voices" and that "we can say with no certainty where one concludes and another begins" (171). We agree with Levenson that no definitive boundaries or certain conclusions can be drawn, but argue that the act of devising a reading—however provisional—is a worthwhile critical act, particularly in a classroom setting.
34. Jewel Spears Brooker and Joseph Bentley, *Reading* The Waste Land*: Modernism and the Limits of Interpretation* (Amherst: University of Massachusetts Press, 1990), 140.
35. Brooke, Hammond and Hirst, "Distinguishing Voices in *The Waste Land* Using Computational Stylistics."
36. Julian Brooke and Graeme Hirst, "A Multi-Dimensional Bayesian Approach to Lexical Style," *Proceedings of the 13th Annual Conference of the North American Chapter of the Association for Computational Lingusitics*, 2013, 673–79.
37. Julian Brooke and Graeme Hirst, "Hybrid Models for Lexical Acquisition of Correlated Styles" Proceedings of the 6th International Joint Conference on Natural Language Processing (2013): 82–90.
38. For details of the Project Gutenberg process, see Brooke, Hammond and Hirst. "Using Models of Lexical Style to Quantify Free Indirect Discourse in Modernist Fiction." *Digital Scholarship in the Humanities* 2.2 (Advance Access, 3 February 2016).
39. Julian Brooke et al., "Unsupervised Multiword Segmentation of Large Corpora Using Prediction-Driven Decomposition of N-Grams" Proceedings of the 25th International Conference on Computational Linguistics (2014): 753–61.
40. Calvin Bedient, *He Do the Police in Different Voices:* The Waste Land *and Its Protagonist* (Chicago: University of Chicago Press, 1986).
41. For a detailed investigation of the "multiform" (140) character of Tiresias, see Jewel Spears Brooker and Joseph Bentley, *Reading* The Waste Land.
42. Among the critics to accuse Eliot of stereotyped representations of female characters are Sandra Gilbert and Susan Gubar, who argue that Eliot "transcribe[s] female language in order to transcend it" (*No Man's Land: The Place of the Woman Writer in the Twentieth Century*, vol. 1: *The War of the Words* [New Haven: Yale University Press, 1988], 236). For more

recent efforts to reassess Eliot's representation of gender, see Cassandra Laity and Nancy Gish, eds., *Gender, Sexuality and Desire in T. S. Eliot* (Cambridge: Cambridge University Press, 2004) and Rachel Potter, "Gender and Obscenity in *The Waste Land*," *The Cambridge Companion to* The Waste Land, ed. Gabrielle McIntire (Cambridge: Cambridge University Press, 2015), 133–46.

43. Adam Hammond and Julian Brooke, "The Brown Stocking: Exploring Voices in Virginia Woolf's To the Lighthouse," 2013, http://brownstocking.org/.
44. See the "What the Computer Said" section of brownstocking.org for details and examples.
45. Graham Pechey, *Mikhail Bakhtin: The Word in the World*, Critics of the Twentieth Century (London ; New York: Routledge, 2007), 208.
46. Pechey, *Mikhail Bakhtin*.
47. Brooke, Hammond, and Hirst, "Using Models of Lexical Style to Quantify Free Indirect Discourse in Modernist Fiction."
48. Majority agreement can produce highly reliable interpretations even when the difficulty of the task results in only moderate inter-annotator agreement. See Beata Beigman Klebanov and Eyal Beigman, "From Annotator Agreement to Noise Models," *Computational Linguistics* 35, no. 4 (2009): 495–503.
49. While a sample of two texts is of course very limited, the fact that FID functions so similarly in two such dissimilar texts—a novel and a short story; an experimental and pervasive employment of the device versus a limited and more conventional one; one text by a female English writer, the other by an Irish male—suggests that the "in-betweenness" of FID will be found to apply more generally.
50. See, for instance, Mary M. Childers, "Virginia Woolf on the Outside Looking Down: Reflections on the Class of Women," *Modern Fiction Studies* 38.1 (Spring 1992): 61–79 and Alison Light, *Mrs. Wolf and the Servants* (New York: Bloomsbury, 2008). More balanced and sympathetic assessments of Woolf's representation of class can be found in Melba Cuddy-Keane, *Virginia Woolf, the Intellectual, and the Public Sphere* (Cambridge: Cambridge University Press, 2003), especially pp. 52–4 and 100–6, and Jean Mills, "Virginia Woolf and the Politics of Class," *A Companion to Virginia Woolf*, ed. Jessica Berman (London: Wiley Blackwell, forthcoming 2016), 219–32.
51. This feature is now publicly available as part of our software package, GutenTag, available at www.projectgutentag.org. See Julian Brooke, Adam Hammond, Graeme Hirst, "GutenTag: an NLP-driven Tool for Digital Humanities Research in the *Project Gutenberg* Corpus," *Workshop on Computational Linguistics for Literature* (North American Association for Computational Linguistics, June 2015): 1–6.