A Tale of Two Cultures:
Bringing Literary Analysis and Computational Linguistics Together

Adam Hammond
Dept of English
University of Toronto
adam.hammond@utoronto.ca

Julian Brooke
Dept of Computer Science
University of Toronto
jbrooke@cs.toronto.edu

Graeme Hirst
Dept of Computer Science
University of Toronto
gh@cs.toronto.edu

Abstract

There are cultural barriers to collaborative effort between literary scholars and computational linguists. In this work, we discuss some of these problems in the context of our ongoing research project, an exploration of free indirect discourse in Virginia Woolf's *To The Lighthouse*, ultimately arguing that the advantages of taking each field out of its "comfort zone" justifies the inherent difficulties.

1 Introduction

Within the field of English literature, there is a growing interest in applying computational techniques, as evidenced by the growth of the Digital Humanities (Siemens et al., 2004). At the same time, a subfield in Computational Linguistics that addresses a range of problems in the genre of literature is gaining momentum (Mani, 2013). Nevertheless, there are significant barriers to true collaborative work between literary and computational researchers. In this paper, we discuss this divide, starting from the classic rift between the two cultures of the humanities and the sciences (Snow, 1959) and then focusing in on a single aspect, the attitude of the two fields towards ambiguity. Next, we introduce our ongoing collaborative project which is an effort to bridge this gap; in particular, our annotation of Virginia Woolf's *To the Lighthouse* for free indirect discourse, i.e. mixtures of objective narration and subjective speech, requires a careful eye to literary detail, and, while novel, interacts in interesting ways with established areas of Computational Linguistics.

2 Background

2.1 The "Two Cultures" Problem

Since the publication of C. P. Snow's influential *The Two Cultures and the Scientific Revolution* (Snow, 1959), the phrase "the two cultures" been used to signify the rift—perceived and generally lamented—between scientific and humanities intellectual cultures. The problem, of course, is the ignorance of each culture with regard to the methods and assumptions of the other, and the resulting impossibility of genuine dialogue between them, preventing them from working together to solve important problems. Many scholars describing the recent rise of the Digital Humanities—the area of research and teaching concerned with the intersection of computing and humanities disciplines—have argued that it effects a reconciliation of the two alienated spheres, bringing scientific methodology to bear on problems within the humanities, many of which had previously been addressed in a less-than-rigorous manner (Hockey, 2004).

From within the discipline of English literature, however, the application of computational methods to literary analysis has frequently been—and continues to be—a matter of considerable controversy (Hoover, 2007; Flanders, 2009). This controversy arises from the perception of many traditional humanists that computational analysis, which aims to resolve dilemmas, seeking singular truth and hard-and-fast answers, is incompatible with the aims of humanistic research, which is often focused on opening up questions for debate rather than resolving them decisively, and often premised on the

Proceedings of the Second Workshop on Computational Linguistics for Literature, pages 1–8,
Atlanta, Georgia, June 14, 2013. ©2013 Association for Computational Linguistics

idea that there are no right answers, only well- and poorly-supported arguments. Critics have responded to these views by arguing that the best computational literary analysis participates in this project of opening up meaning, arguing that it is not a rejection of literary reading but rather a method for carrying it out more efficiently and extending it to more texts (Ramsay, 2007), and that computational modelling, even when unsuccessful, allows for the application of the scientific method and thus carries the potential for intellectual advancement not possible with purely anecdotal evidence (McCarty, 2005). Despite such counter-arguments, however, the fear remains widespread among traditional literary scholars that the rise of computational analysis will entail the loss of certain sacred assumptions of humanistic inquiry.

2.2 Ambiguity Across the "Cultures"

We argue, though, that these fears are not without basis, particularly when one considers the very different approaches to the question of ambiguity in the two specific disciplines involved in our project: English Literature and Computational Linguistics. Here, the rift of the two cultures remains evident.

A major focus of literary scholarship since the early twentieth century has been the semantic multiplicity of literary language. Such scholarship has argued that literature, distinct from other forms of discourse, may be deliberately ambiguous or polysemous and that literary analysis, distinct from other analytic schools, should thus aim not to resolve ambiguity but to describe and explore it. This was a central insight of the early twentieth-century school, the New Criticism, advanced in such works as William Empson's *Seven Types of Ambiguity* (Empson, 1930) and Cleanth Brooks's *The Well Wrought Urn* (Brooks, 1947), which presented ambiguity and paradox not as faults of style but as important poetic devices. New Criticism laid out a method of literary analysis centred on the explication of the complex tensions created by ambiguity and paradox, without any effort to resolve them. Also in the first half of the twentieth century, but independently, the Russian critic Mikhail Bakhtin developed his theory of dialogism, which valorized "double-voiced" or polyphonic works that introduce multiple, competing perspectives—particularly voices—that present conflicting ideologies (Bakhtin, 1981). Bakhtin, who wrote his seminal work "Discourse in the Novel" under a Stalinist sentence of exile, particularly valued works that enacted the free competition of ideologically opposed voices. In a similar spirit, but independently of Bakhtin, the German critic Erich Auerbach described the "multi-personal representation of consciousness", a narrative technique in which the writer, typically the narrator of objective facts, is pushed entirely into the background and the story proceeds by reflecting the individual consciousnesses of the characters; Auerbach argued that this was a defining quality of modernist (early twentieth-century) literature (Auerbach, 1953). In the second half of the twentieth century, this critical emphasis on ambiguity and paradox developed in an extreme form into the school of deconstructive criticism, which held a theory of the linguistic sign according to which determinate linguistic meaning is considered logically impossible. Deconstructive literary analysis proceeds by seeking out internal contradictions in literary texts to support its theory of infinitely ambiguous signification.

In Computational Linguistics, by contrast, ambiguity is almost uniformly treated as a problem to be solved; the focus is on disambiguation, with the assumption that one true, correct interpretation exists. In the sphere of annotation, for instance, there is an expectation that agreement between annotators, as measured by statistics such as kappa (Di Eugenio and Glass, 2004), reach levels (generally 0.67 or higher) where disagreements can be reasonably dismissed as noise; the implicit assumption here is that subjectivity is something to be minimized. The challenge of dealing with subjectivity in CL has been noted (Alm, 2011), and indeed there are rare examples in the field where multiple interpretations have been considered during evaluations—for instance, work in lexical cohesion (Morris and Hirst, 2005) and in using annotator disagreements as an indicator that two words are of similar orientation (Taboada et al., 2011)—but they are the exception. Work in CL focused on literary texts tends towards aspects of the texts which readers would not find particularly ambiguous, for example identifying major narrative threads (Wallace, 2012) or distinguishing author gender (Luyckx et al., 2006).

3 A Collaborative Research Agenda

The obvious solution to the problem of the "two cultures"—and one that has often been proposed (Friedlander, 2009)—is interdisciplinary collaboration. But while there are many computational linguists working in literary topics such as genre, and many literary scholars performing computational analysis of literature, genuine collaboration between the disciplines remains quite rare. Over the past two years, we have undertaken two collaborative projects—one mostly complete, one ongoing—which aim at such genuine collaboration, and in so doing seek to bridge the real rift between scientific and humanities cultures.[1] Each of these projects is multi-faceted, seeking (a) to produce meaningful research within both disciplines of Computational Linguistics and English Literature; (b) to provide educational experience which broadens the disciplinary horizons of the undergraduate students involved in the projects; and (c) to provide a model of collaborative research that will spur further such "culture-spanning" projects.

Each of our projects was launched in the context of a course entitled "The Digital Text" offered by the Department of English at the University of Toronto. The first author, whose background is in English Literature, is instructor of the course, while the second author, a graduate student in Computer Science, was assigned as a teaching assistant. Working together with the third author, we have designed these projects collaboratively.

The first project, which we call "He Do the Police in Different Voices",[2] was carried out in 2011–12 (Hammond, 2013). Focused on a "multi-personal" poem, *The Waste Land* (1922) by T.S. Eliot, it encompassed each of the three aspects of our projects outlined above; in particular, it was motivated by a research question of interest to both disciplines: could we identify the points in *The Waste Land* where the style changes, where one "voice" gives way to another? A computational approach promised to bring added rigor as well as a degree of objectivity to this question, which humanities methods had proven unable to resolve in almost a century of debate. Both because poetry is dense in signification, and because the multiple voices in *The Waste Land* are a deliberate effect achieved by a single author rather than a disguised piecing together of the works of multiple authors, the question provided a meaningful challenge to the computational approach, an unsupervised vector-space model which first segments by identifying points of stylistic change (Brooke et al., 2012) and then clusters the resulting segments together into voices (Brooke et al., 2013).

This research project was tightly integrated into the curriculum of "The Digital Text". Students were instructed in the use of the Text Encoding Initiative (TEI) XML guidelines,[3] and each of the students provided one annotation related to voice as part of a marked assignment. Students also participated in an online poll in which they indicated every instance in which they perceived a vocal switch in the poem, and their responses were used in the construction of a gold standard for the evaluation of our computational approach.

Once they were complete, we developed our results into a publicly accessible website.[4] This website promises to encourage collaboration between literary scholars and computational linguists by explaining the project and our results in language accessible to both, and by producing a new digital edition of the poem based on our findings. Human and computer readings of the poem are presented side-by-side on the website, to demonstrate that each interprets the poem in different ways, but that neither of these methods is absolutely valid. Rather, we encourage website visitors to decide for themselves where they believe that the vocal switches occur, and we provide an interactive interface for dividing the poem up according to their own interpretation. In addition to serving as a model of collaboration between English Literature and Computational Linguistics—and also serving as a teaching tool for instructors of *The Waste Land* at any level—the site is thus useful to us as a source of further data.

[1] In addition, the third author was part of a separate collaborative project between our departments (Le et al., 2011), though the aim of that project was not literary analysis.

[2] This is a reference to Eliot's working title for *The Waste Land*, which in itself is a reference to a talented storyteller in *Our Mutual Friend* by Charles Dickens; another Dickens novel is alluded to in the title of this paper.

[3] http://www.tei-c.org/Guidelines/

[4] http://www.hedothepolice.org

4 The "Brown Stocking" Project

4.1 Free Indirect Discourse in *To the Lighthouse*

Our second, ongoing project, "The Brown Stocking", focuses on a literary text deliberately chosen for its deeply ambiguous, polysemous, dialogic nature: Virginia Woolf's (1927) *To the Lighthouse* (*TTL*). Woolf's novel was produced at the same time that critical theories of ambiguity and polyvocality were being developed, and indeed was taken as a central example by many critics. Our project takes its title from the final chapter of Erich Auerbach's *Mimesis*, in which Auerbach presents *TTL* as the representative text of modernist literature's "multipersonal representation of consciousness" (Auerbach, 1953). For Auerbach, there are two principal distinguishing features in Woolf's narrative style. The first is the tendency, already noted, to "reflect" incidents through the subjective perspectives of characters rather than presenting them from the objective viewpoint of the author; thus *TTL* becomes a work in which there is more than one order and interpretation. Woolf's technique not only introduces multiple interpretations, however, but also blurs the transitions between individual perspectives, making it difficult to know in many instances who is speaking or thinking.

Woolf achieves this double effect—multiple subjective impressions combined with obscuring of the lines separating them from the narrator and from one another—chiefly through the narrative technique of free indirect discourse (also known as free indirect style). Whereas direct discourse reports the actual words or thoughts of a character, and indirect discourse summarizes the thoughts or words of a character in the words of the entity reporting them, free indirect discourse (FID) is a mixture of narrative and direct discourse (Abrams, 1999). As in indirect discourse, the narrator employs third-person pronouns, but unlike indirect discourse, the narrator includes words and expressions that indicate subjective or personalized aspects clearly distinct from the narrator's style. For example, in the opening sentences of *TTL*:

> "Yes, of course, if it's fine tomorrow," said Mrs. Ramsay. "But you'll have to be up with the lark," she added. To her son these words conveyed an extraordinary joy, as if it were settled, the expedition were bound to take place, and the wonder to which he had looked forward, for years and years it seemed, was, after a night's darkness and a day's sail, within touch.

we are presented with two spans of objective narration (*said Mrs. Ramsay* and *she added*) and two passages of direct discourse, in which the narrator introduces the actual words of Mrs. Ramsay (*"Yes, of course, if it's fine tomorrow"* and *"But you'll have to be up with the lark"*). The rest of the passage is presented in FID, mixing together the voices of the narrator, Mrs. Ramsay, and her son James: while the use of third-person pronouns and the past tense and clearly indicates the voice of the narrator, phrases such as *for years and years it seemed* clearly present a subjective perspective.

In FID's mixing of voices, an element of uncertainty is inevitably present. While we can be confident of the identity of the voice speaking certain words, it remains unclear whether other words belong to the narrator or a character; in this case, it is not clear whether *for years and years it seemed* presents James's actual thoughts, Mrs. Ramsay's summary of her son's thoughts, the narrator's summary of James's thoughts, the narrator's summary of Mrs. Ramsay's summary of James's thoughts, etc. Abrams (1999) emphasizes uncertainty as a defining trait of FID: the term "refers to the way, in many narratives, that the reports of what a character says and thinks shift in pronouns, adverbs, and grammatical mode, as we move—or sometimes hover—between the direct narrated reproductions of these events as they occur to the character and the indirect representation of such events by the narrator". FID, with its uncertain "hovering", is used throughout *TTL*; it is the principal technical means by which Woolf produces ambiguity, dialogism, and polysemy in the text. It is thus the central focus of our project.

In Literary Studies, Toolan (2008) was perhaps the first to discuss the possibility of automatic recognition of FID, but his work was limited to a very small, very informal experiment using a few *a priori* features, with no implementation or quantitative analysis of the results. Though we are not aware of work in Computational Linguistics that deals with this kind of subjectivity in literature—FID is included in the narrative annotation schema of Mani (2013), but it is not given any particular attention within that framework—there are obvious connections with sentence-level subjectivity analysis (Wilson et al., 2005) and various other stylistic tasks, including authorship profiling (Argamon et al., 2007). Since the subjective nature of these passages is often expressed through specific lexical choice, it would be interesting to see if sentiment dictionaries (Taboada et al., 2011) or other stylistic lexical resources such as dictionaries of lexical formality (Brooke et al., 2010) could be useful.

4.2 Our Approach

Our project is proceeding in four stages: an initial round of student annotation, a second round of student annotation, computational analysis of these annotations, and the development of a project website. In the first stage, we had 160 students mark up a passage of between 100–150 words in accordance with TEI guidelines. Students were instructed to use the TEI `said` element to enclose any instance of character speech, to identify the character whose speech is being introduced, and to classify each of these instances as either direct, indirect, or free indirect discourse and as either spoken aloud or thought silently. Because there are often several valid ways of interpreting a given passage, and because we are interested in how different students respond to the same passage, each 100–150 word span was assigned to three or four students. This first round of annotation focused only on the first four chapters of *TTL*. Raw average agreement of the various annotations at the level of the word was slightly less than 70%,[5] and though we hope to do better in our second round, levels of agreement typically required are likely to be beyond our reach due to the nature of the task. For example, all four sudents responsible for the passage cited above agreed on the tagging of the first two sentences; however, two students read the third sentence as FID mixing the voices of the narrator and Mrs. Ramsay, and two read it as FID mixing the voice of the narrator and James. Though they disagree, these are both valid interpretations of the passage.

In the second round of annotation, with 160 different student annotators assigned slightly longer spans of 200–300 words, we are focusing on the final seven chapters of *TTL*. We have made several minor changes to our annotation guidelines, and two significant changes. First, we now ask that in every span of text which students identify as FID, they explicitly identify the words that they regard as clearly coming from the subjective perspective of the character. We believe this will help students make a valid, defensible annotation, and it may also help with the computational analysis to follow. Second, we are also allowing embedded tags, for instances of direct or indirect discourse within spans of FID, which were confusing to students in the initial round. For instance, students would now be able to tag the above-cited passage of as a span of FID mixing the narrator's and Mrs. Ramsay's words, inside of which Mrs. Ramsay introduces an indirect-discourse rendering of her son's thoughts. Moving from a flat to a recursive representation will naturally result in additional complexity, but we believe it is necessary to capture what is happening in the text.

Once this second round of tagging is complete, we will begin our computational analysis. The aim is to see whether we can use supervised machine learning to replicate the way that second-year students enrolled in a rigorous English literature program respond to a highly complex text such as *TTL*. We are interested to see whether the subjective, messy data of the students can be used to train a useful model, even if it is inadequate as a gold standard. If successful, this algorithm could be deployed on the remaining, untagged sections of *TTL* (i.e. everything between the first four and last seven chapters) and produce meaningful readings of the text. It would proceed by (a) identifying passages of FID (that is, passages in which it is unclear whether a particular word belongs to the narrator or a character); (b) making an interpretation of that passage (hypothesizing as to which particular voices are being mixed); and (c) judging the likely validity of this interpretation. It would seek not only to *identify* spans of vocal ambiguity, but also to *describe* them, as far as possible. It would thus not aim strictly at disambiguation—at producing a right-or-wrong

[5] Since each passage was tagged by a different set of students, we cannot apply traditional kappa measures. Raw agreement overestimates success, since unlike kappa it does not discount random agreement, which in this case varies widely across the different kinds of annotation.

reading of the text—but rather at producing the best possible interpretation. The readings thus generated could then be reviewed by an independent expert as a form of evaluation.

Finally, we will develop an interactive website for the project. It will describe the background and aims of the project, present the results from the first three stages of the project, and also include an interface allowing visitors to the site to annotate the text for the same features as the students (via a Javascript interface, i.e. without having to manipulate the XML markup directly). This will provide further annotation data for our project, as well as giving instructors in English Literature and Digital Humanities a resource to use in their teaching.

5 Discussion

We believe our approach has numerous benefits on both sides of the divide. From a research perspective, the inter-disciplinary approach forces participants from both English Literature and Computational Linguistics to reconsider some of their fundamental disciplinary assumptions. The project takes humanities literary scholarship out of its "comfort zone" by introducing alien and unfamiliar methodologies such as machine learning, as well as by its basic premise that FID—by definition, a moment of uncertainty where the question of who is speaking is unresolved—can be detected automatically. Even though many of these problems can be linked with classic Computational Linguistics research areas, the project likewise takes Computational Linguistics out of its comfort zone by seeking not to resolve ambiguity but rather to identify it and, as far as possible, describe it. It presents an opportunity for a computational approach to take into account a primary insight of twentieth-century literary scholarship: that ambiguity and subjectivity are often desirable, intentional qualities of literary language, not problems to be solved. It promises literary scholarship a method for extending time-consuming, laborious human literary readings very rapidly to a vast number of literary texts, the possible applications of which are unclear at this early stage, but are surely great.

While many current major projects in computer-assisted literary analysis operate on a "big-data" model, drawing conclusions from analysis of vast numbers of lightly annotated texts, we see advantages in our own method of beginning with a few heavily-annotated texts and working outward. Traditional literary scholars often object that "big-data" readings take little or no account of subjective, human responses to literary texts; likewise, they find the broad conclusions of such projects (that the nineteenth century novel moves from telling to showing (Heuser and Le-Khac, 2012); that Austen is more influential than Dickens (Jockers, 2012)) difficult to test (or reconcile) with traditional literary scholarship. The specific method we are pursuing—taking a great number of individual human readings of a complex literary text and using them as the basis for developing a general understanding of how FID works—promises to move literary analysis beyond merely "subjective" readings without, however, denying the basis of all literary reading in individual, subjective responses. Our method indeed approaches the condition of a multi-voiced modernist literary work like *TTL*, in which, as Erich Auerbach perceived, "overlapping, complementing, and contradiction yield something that we might call a synthesized cosmic view". We too are building our synthetic understanding out of the diverse, often contradictory, responses of individual human readers.

Developing this project in an educational context—basing our project on readings developed by students as part of marked assignments for "The Digital Text"—is likewise beneficial to both cultures. It forces humanities undergraduates out of their comfort zone by asking them to turn their individual close readings of the text into an explicit, machine-readable representation (in this case, XML). Recognizing the importance of a sharable language for expressing literary features in machine-readable way, we have employed the standard TEI guidelines mark-up with as few customizations as possible, rather than developing our own annotation language from the ground up. The assignment asks students, however, to reflect critically on whether such explicit languages can ever adequately capture the polyvalent structures of meaning in literary texts; that is, whether there will always necessarily be possibilities that can't be captured in the tag set, and whether, as such, an algorithmic process can ever really "read" literature in a useful way. At the same time, this method has potentially great benefits to the development of such algorithmic readings, precisely by making available machine-readable approximations of how readers belonging to another "culture"—humanities undergraduates—respond to a challenging literary text. Such annotations would not be possible from a pool of annotators trained in the sciences, but could only come from students of the humanities with a basic understanding of XML. We do not believe, for example, workers on Amazon Mechanical Turk could reliably be used for this purpose, though it might be interesting to compare our 'studentsourcing' with traditional crowdsourcing techniques.

Our approach also faces several important challenges. Certainly the largest is whether an algorithmic criticism can be developed that could come to terms with ambiguity. The discipline of literary studies has long taught its students to accept what the poet John Keats called "negative capability, that is, when a man is capable of being in uncertainties, mysteries, doubts, without any irritable searching after fact and reason" (Keats, 2002). Computational analysis may simply be too fundamentally premised on "irritable searching after fact and reason" to be capable of "existing in uncertainty" in the manner of many human literary readers. Even if we are able to develop a successful algorithmic method of detecting FID in Woolf, this method may not prove applicable to other literary texts, which may employ the device in highly individual manners; *TTL* may prove simply too complex—and employ too much FID—to serve as a representative sample text. At a more practical level, even trained literature students do not produce perfect annotations: they make errors both in XML syntax and in their literary interpretation of *TTL*, a text that proves elusive even for some specialists. Since we do not want our algorithm to base its readings on invalid student readings (for instance, readings that attribute speech to a character clearly not involved in the scene), we face the challenge of weeding out bad student readings—and we will face the same challenge once readings begin to be submitted by visitors to the website. These diverse readings do, however, also present an interesting possibility, which we did not originally foresee: the development of a reader-response "map" showing how human readers actually interpret (and in many cases misinterpret) complex modernist texts like *TTL*.

6 Conclusion

Despite the philosophical and technical challenges that face researchers in this growing multidisciplinary area, we are increasingly optimistic that collaboration between computational and literary researchers is not only possible, but highly desirable. Interesting phenomena such as FID, this surprising melding of objective and personal perspective that is the subject of the current project, requires experts in both fields working together to identify, annotate, and ultimately model. Though fully resolving the rift between our two cultures is not, perhaps, a feasible goal, we argue that even this early and tentative collaboration has demonstrated the potential benefits on both sides.

Acknowledgements

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

References

M. H. Abrams. 1999. *A Glossary of Literary Terms*. Harcourt Brace, Toronto, 7th edition.

Cecilia Ovesdotter Alm. 2011. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112.

Shlomo Argamon, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan. 2007. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 7:91–109.

Erich Auerbach. 1953. *Mimesis: The Representation of Reality in Western Literature*. Princeton University Press, Princeton, NJ.

Mikhail Mikhailovich Bakhtin. 1981. Discourse in the novel. In Michael Holquist, editor, *The Dialogic Imagination: Four Essays*, pages 259–422. Austin: Univeristy of Texas Press.

Julian Brooke, Tong Wang, and Graeme Hirst. 2010. Automatic acquisition of lexical formality. In *Proceed-*

ings of the 23rd International Conference on Computational Linguistics (COLING '10), Beijing.

Julian Brooke, Adam Hammond, and Graeme Hirst. 2012. Unsupervised stylistic segmentation of poetry with change curves and extrinsic features. In *Proceedings of the 1st Workshop on Computational Literature for Literature (CLFL '12)*, Montreal.

Julian Brooke, Graeme Hirst, and Adam Hammond. 2013. Clustering voices in *the Waste Land*. In *Proceedings of the 2nd Workshop on Computational Literature for Literature (CLFL '13)*, Atlanta.

Cleanth Brooks. 1947. *The Well Wrought Urn*. Harcourt Brace, New York.

Barbara Di Eugenio and Michael Glass. 2004. The kappa statistic: a second look. *Computational Linguistics*, 30(1):95–101, March.

T.S. Eliot. 1971. The Waste Land. In *The Complete Poems and Plays, 1909–1950*, pages 37–55. Harcourt Brace Jovanovich, New York.

William Empson. 1930. *Seven Types of Ambiguity*. Chatto and Windus, London.

Julia Flanders. 2009. Data and wisdom: Electronic editing and the quantification of knowledge. *Literary and Linguistic Computing*, 24(1):53–62.

Amy Friedlander. 2009. Asking questions and building a research agenda for digital scholarship. Working Together or Apart: Promoting the Next Generation of Digital Scholarship. Report of a Workshop Cosponsored by the Council on Library and Information Resources and The National Endowment for the Humanities, March.

Adam Hammond. 2013. He do the police in different voices: Looking for voices in *The Waste Land*. Seminar: "Mapping the Fictional Voice" American Comparative Literature Association (ACLA).

Ryan Heuser and Long Le-Khac. 2012. A quantitative literary history of 2,958 nineteenth-century British novels: The semantic cohort method. Stanford Literary Lab Pamphlet No. 4. http://litlab.stanford.edu/LiteraryLabPamphlet4.pdf .

Susan Hockey. 2004. The history of humanities computing. In Ray Siemens, Susan Schreibman, and John Unsworth, editors, *A Companion to Digital Humanities*. Blackwell, Oxford.

David L. Hoover. 2007. Quantitative analysis and literary studies. In Ray Siemens and Susan Schreibman, editors, *A Companion to Digital Literary Studies*. Blackwell, Oxford.

Matthew L. Jockers. 2012. Computing and visualizing the 19th-century literary genome. Presented at the Digital Humanities Conference. Hamburg.

John Keats. 2002. *Selected Letters*. Oxford University Press, Oxford.

Xuan Le, Ian Lancashire, Graeme Hirst, and Regina Jokel. 2011. Longitudinal detection of dementia through lexical and syntactic changes in writing: A case study of three British novelists. *Literary and Linguistic Computing*, 26(4):435–461.

Kim Luyckx, Walter Daelemans, and Edward Vanhoutte. 2006. Stylogenetics: Clustering-based stylistic analysis of literary corpora. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC '06)*, Genoa, Italy.

Inderjeet Mani. 2013. *Computational Modeling of Narrative*. Morgan & Claypool.

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

Jane Morris and Graeme Hirst. 2005. The subjectivity of lexical cohesion in text. In James G. Shanahan, Yan Qu, and Janyce M. Wiebe, editors, *Computing Attitude and Affect in Text*. Springer, Dordrecht, The Netherlands.

Stephen Ramsay. 2007. Algorithmic criticism. In Ray Siemens and Susan Schreibman, editors, *A Companion to Digital Literary Studies*. Blackwell, Oxford.

Ray Siemens, Susan Schreibman, and John Unsworth, editors. 2004. *A Companion to Digital Humanities*. Blackwell, Oxford.

C. P. Snow. 1959. *The Two Cultures and the Scientific Revolution*. Cambridge University Press, New York.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Michael Toolan. 2008. Narrative progression in the short story: First steps in a corpus stylistic approach. *Narrative*, 16(2):105–120.

Byron Wallace. 2012. Multiple narrative disentanglement: Unraveling *Infinite Jest*. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1–10, Montréal, Canada, June. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT/EMNLP '05, pages 347–354.

Virginia Woolf. 1927. *To the Lighthouse*. Hogarth, London.